BUILDING A SYSTEMATIC ANALYTIC PIPELINE – BIG DATA INNOVATION IN HEALTHCARE

A Thesis

Presented to

The Academic Faculty

by

Yuanbo Wang

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy in the

College of Biological Sciences

Georgia Institute of Technology

December 2019

BUILDING A SYSTEMATIC ANALYTIC PIPELINE – BIG DATA INNOVATION IN HEALTHCARE

Approved by:

Professor Eva K. Lee,
Committee Chair
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor King Jordan
College of Biological Sciences
Georgia Institute of Technology

Professor Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Frederik Vannberg
School of Biological Sciences
Georgia Institute of Technology

Professor Alfred Merrill
School of Biological Sciences
Georgia Institute of Technology

Dr. Shatavia Morrison
Center for Disease Control and Prevention

Date Approved: 08/15/2019

This thesis is dedicated to my parents, Jiawang Wang and Fang Wang, and my wife, Hanyan Li.

ACKNOWLEDGEMENTS

First, I would like to offer my advisor, Dr. Eva Lee, my deepest gratitude. Since I started my PhD program at Georgia Tech, I have felt extremely fortunate to have the opportunity to work with her on health science projects which have significant impact on the entire society. To me, Dr. Lee is not only a knowledgeable instructor and mentor, but also an inspirational role model who works hard and enjoys challenges. She would always tell me that the reason I was tasked with challenging work is because she believed I had the ability to do it. These words always motivated me at stressful times and pushed me to become a determined person today.

Second, I would like to thank each member of my committee. Dr. King Jordan taught my first ever bioinformatics class, during which I learned the knowledge and tools that would lay the foundations for my dissertation research. His classes also provided valuable hands-on opportunities in subjects including functional annotation and gene prediction. Dr. Yajun Mei identified some of the missing components of my thesis and provided valuable ideas on how to improve its technical depth and scientific rigorousness. Dr. Fred Vannberg showed me ways to improve the practicality of my thesis work and enriched its scope. Dr. Shatavia Morrison has mentored me since my fellowship at the CDC four years ago and provided me with both academic and professional advice. Dr. Merrill's Cancer Biology class was perhaps one of the most informative class I have taken at Georgia Tech and a reminder of the potential life-changing significance of biomedical research. Without the support of my committee members, I would not have been able to reach my fullest potential in completing my thesis work.

Third, I would like to thank a few professors and staffs whom I met during my PhD studies. In particular, Dr. Randall Guensler provided me with opportunities to work on open-ended transportation-related projects that not only trained my problem-solving skills but also motivated me to bridge the gap between transportation and public health with my interdisciplinary knowledge. Lisa Redding was my academic program coordinator throughout these five years and was most devoted to helping her students. She would strive to provide us the best support outside of academic research and has always been efficient and reliable.

Finally, I want to thank my family and friends. I would not have been here if it were not for my mom and dad. Although they are thousand miles away from me, they have never failed to provide the love and support I needed. My grandmother, aunts, and uncles would keep motivating me to pursue my goals, especially at times when I feel lost. My wife, Ann, who will be earning her PhD degree as well this week, has been the most caring and inspiring companion I could ever ask for. We stood by each other throughout our journey together, overcoming many difficulties which I would not have been able to face alone. My friends, whether old and new, have kept me accompanied at times of needs. In particular, Matthew Hagen, who graduated from Dr. Lee's lab four years ago, has given me many valuable advices throughout my PhD studies. For other family members and friends whose name I did not mention here, but have always had my back, I have not forgotten what you have done for me, and will do my best to not let you down.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Electronic Health Records (EHR) containing large amount of patient data present both opportunities and challenges to industry, policy makers, and researchers. Data-driven healthcare utilizing big data in EHR has the potential to revolutionize care delivery while reducing costs. However, for researchers, policymakers, and practitioners to take full advantage of the benefits that electronic records can provide, several challenges must be addressed: 1) Extraction and coding methods for EHR data must be strategically designed to address issues of data quantity, quality, and patient confidentiality; 2) Standardization of clinical terminologies is essential in facilitating interoperability among EHR systems and allows for multi-site comparative effectiveness studies; 3) Effective methods for mining longitudinal health data common in the EHR are critical for revealing disease progression, treatment patterns, and patient similarities, all of which play important role in clinical decision support and treatment improvement; 4) Advanced machine learning techniques are necessary for early detection and prognosis of disease and identifying critical factors that impact patient outcome and; 5) Practical intervention strategies must be developed to address healthcare disparity in rural and remote areas with lack of resources and access. This thesis focuses on these five issues by developing a systematic analytic pipeline for big data in healthcare. Specifically, innovative strategies are developed for information extraction, clinical terminology mapping, time-series mining and clustering, feature selection and discriminatory modeling. Finally, practical implementation methods for telehealth services are designed to reduce healthcare disparity in underserved rural Appalachian counties in Georgia.

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

Electronic health record (EHR) plays an important role in advancing clinical and operational processes. Although early clinical medical records first appeared in 1600 BC, it was not until 1900 that it was put into regular use [1]. The launch of the 10-year-effort to create a national electronic medical record system by the United State government in 2004 helped fuel its rapid adoption and medical advance [2]. As of 2015, 80 percent of U.S. hospitals had adopted a basic electronic health record keeping system [3]. The value of EHR data is increasingly recognized by healthcare organization and government. Its utilizations significantly changed the patient-clinic interaction process [1].

For researchers, policymakers, and practitioners to take full advantage of the benefits that electronic records can provide, critical preparation steps including extracting relevant information from the EHR database [4], de-identifying and encrypting Protected Health Information (PHI) [5], and standardizing heterogeneous data [6] must be optimized. Analysis of big data require advanced Artificial Intelligence. Machine learning algorithms applied to big data in healthcare can identify favorable actions, predict risks, mortality, and length of stay. It is also part of a growing trend towards personalized, predictive medicine. EHR also offers access to longitudinal data including laboratory tests, prescriptions, and vital signs recorded during the delivery care, which can provide insights to disease progression and treatment outcome. These data records form unevenly spaced time series, rendering them difficult to analyze and understand with existing algorithms.

Knowledge discovered from clinical data should be effectively transformed into practical implementations for multiple sites. Despite advances in communication technologies, many rural and remote areas still lack healthcare staff and services. Telehealth helps eliminate distance barriers and improve access to medical services that would often not be consistently available in distant rural communities. However, due to compounding challenges in confidentiality, technical difficulties, physician licensing, and reimbursement, it remains a promising but underdeveloped area of service.

1.2 CONTRIBUTION OF THIS THESIS

The work presented in this thesis makes several contributions to information extraction, knowledge discovery, and practical implementation strategies using EHR and other clinical big data:

- An information extraction and concept mining framework for structured and unstructured data within the EHR is developed. Highly efficient queries are designed to identify and extract datasets with comprehensive coverage of target patient cohorts. An end-to-end pipeline is developed for extracting and consolidating key clinical features from unstructured narrative documents.
- Multi-site heterogeneous EHR data are standardized via an automated mapping system, which links raw phrases and codes recorded in various clinical terminology and coding systems to Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED-CT) concepts.
- Techniques are introduced to extract knowledge from longitudinal clinical data in the form of time series. Specifically, univariate and multivariate time series clustering approaches are developed to characterize patient disease severity and recovery progress using laboratory and vitals recorded during delivery of care.

- Discriminatory features that inform optimized decisions during the treatment of diabetes, prostate cancer, and cardiovascular disease are identified with various state-of-art supervised learning algorithms.
- A p-median facility location problem is formulated to set up telehealth sites to efficiently serve the rural communities. A prototypical design of a low-cost remote patient monitoring device that supports remote care management, secure communication, and monitoring of disease conditions is developed and its feasibility is evaluated using a Return of Investment (ROI) analysis.

1.3 STRUCTURE OF THIS THESIS

Chapter 2 provides the background knowledge of many existing technologies that provide the foundation for this dissertation. The concept of Named Entity Recognition (NER) and Clinical Terminology Systems, various distance metrics and clustering algorithms for time series, and supervised learning techniques are introduced. Challenges associated with healthcare disparity are presented, and the significance of this work is described.

Chapter 3 introduces the design and implementation of an efficient and comprehensive data extraction methods from epic EHR database. By designing efficient queries to extract data covering all aspects of patient information (i.e. demographics, diagnosis, laboratory and surgical procedures, medications, clinical notes, etc.), and developing a pipeline for identifying and extracting key clinical concepts from unstructured narrative notes, big data are explored with a holistic approach.

Chapter 4 develops an effective concept standardization process among the multitude of available clinical terminologies to facilitate sharing and exchange of healthcare information. An

automated pipeline is created to streamline the mapping process so that it can be easily adapted and applied to other EHR systems and establish data interoperability for quality care delivery and coordination among multiple healthcare sites.

Chapter 5 advances unsupervised learning algorithms for characterizing patient treatment outcomes based on longitudinal data. Challenges associated with mining laboratory and vitals data are presented and addressed. By utilizing existing similarity metrics and clustering algorithms, a new approach is developed to cluster irregular (unevenly spaced, with unequal lengths) multivariate time series (MTS)..

Chapter 6 compares and advances supervised learning methods to predict treatment outcome and identify discriminatory features for clinical practice and outcome improvement. Challenges associated with feature representation, feature selection, and imbalanced data in machine learning tasks are addressed.

Chapter 7 explores current status of and challenges in telehealth and design intervention methods to improve rural health access. A p-median facility location problem is formulated to identify optimal locations for setting up telehealth sites to efficiently serve the rural communities. A low-cost remote patient monitoring device that supports remote care management, secure communication, and monitoring of disease conditions is designed, and its practicality is evaluated through a Return of Investment (ROI) analysis.

CHAPTER II

BACKGROUND AND SIGNIFICANCE

Many information extraction techniques have been applied to the clinical domain. These techniques have developed from simple pattern matching, semantic-based, to more complex machine learning-based approaches. However, understanding unstructured clinical text remains a big challenge. Various clinical terminology systems have been developed to address the issue of "Variety"—one of the four "V"s that define big data. Both integrative and specialized terminology systems play important roles in promoting the interoperability of clinical data.

Knowledge learned from volumes of EHR data have proven valuable in many ways—optimizing processes, personalizing treatment, informing decisions, improving patient experiences, and reducing costs. Both supervised and unsupervised machine learning algorithms have great utility in analysis of EHR.

Practical implementation guidelines and strategies have been developed to address healthcare disparity but have met many hurdles due to the sensitivity of protected patient health information, technical difficulties, and policy issues. Telehealth services and remote patient monitoring technologies have started to transform traditional healthcare delivery and improve patients' quality of life in underserved areas.

This chapter covers established techniques and prominent challenges in each of these areas and identifies many gaps that need to be addressed.

2.1 INFORMATION EXTRACTION AND PHI ENCRYPTION

It is challenging to establish an efficient data extraction schema for EHR due to the complexity of data and lack of data standards. A common task in EHR is case detection – identifying a cohort of

patients with a certain condition or symptom. Coded data such as International Classification of Disease (ICD) codes are often not sufficient or accurate [7, 8]. Informatics approaches combining structured EHR data with narrative text data achieve better performance [9, 10]. Key clinical items can be extracted from narrative texts with simple methods such as pattern matching using regular expression [11-13], full or partial parsing based on morpho-semantems [14, 15], and syntactic and semantic analysis [16, 17]. Recently, more complex statistical and rule-based machine learning approaches [18, 19] have been developed to tackle this challenge. Biomedical Named Entity Recognition (NER) – the "task of identifying words and phrases in free text that belong to certain classes of interest" [20], allows users to identify key clinical concepts such as physician visits, referrals, dietary management, and suspected problems normally not present in structured data tables. Three main approaches have been used to accomplish this task: rule / lexicon-based [21, 22], supervised learning [23-25], and hybrid Long-Short Term Memory (LSTM) – Conditional Random Fields (CRF) [26, 27].

Negation detection, which identifies the negative sense of a concept, is another essential task accompanying NER, since the presence of negations can yield false-positive detections because medical personnel are trained to include pertinent negatives in their reports [28]. It has been achieved through rule-based / syntactic parsing [29-31] and machine learning [32-34] approaches.

Once patient information is extracted, data security and confidentiality must be ensured through de-identification steps. According to Health Insurance Portability and Accountability Act (HIPAA), patients' Protected Health Information (PHI) must be de-identified or anonymized for commercial and research interest. PHI exists in both structured and unstructured clinical records [35]. This includes patient names, addresses, phone numbers, etc. Manual and rule-based or

lexicon-based methods have been used to achieve PHI de-identification [36-38], but they are extremely time-consuming and can be inaccurate. Machine learning approaches have also been developed [39-41]. However, due to the complexity of data schemas and the heterogeneity of data structures, it is very challenging to detect PHI with high sensitivity.

2.2 CLINICAL DATA INTEROPERABILITY AND TERMINOLOGY MAPPING

Data from EHR and other forms of clinical and biomedical data can reveal critical variables that impact treatment outcomes and inform allocation of limited time and resources, allowing physicians to provide evidence-based treatment plans that suit the needs of individual patient. On a larger scale, realistically modifiable social determinants of health that will improve community health can potentially be discovered and addressed. Tackling the problem of data heterogeneity is essential for conducting predictive analytics using artificial intelligence. Many clinical records in the EHR adhere to different terminology systems and can cause problems such as data redundancy and inconsistency, exacerbating the issue of "curse of dimensionality" associated with big data, and hindering the performance of automated machine learning models.

Standardized clinical terminologies are essential in facilitating interoperability among EHR systems. They allow seamless sharing and exchange of healthcare information for quality care delivery and coordination among multiple sites. However, the volume and number of available clinical terminologies are large and are expanding. Due to the increase in medical knowledge, and the continued development of more advanced computer systems, the usage of medical terminologies has extended beyond diagnostic classification [3]. The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) is the most comprehensive healthcare terminology and is widely implemented and used in clinical settings [42-47]. It provides foundation for patient-practitioner interactions, laboratory testing and reporting, clinical problems,

pharmaceutical ingredients, medical and surgical procedures [48]. When complemented with other vocabularies (e.g. LOINC), SNOMED-CT can be used to build the enterprise reference terminology of a healthcare organization [48]. The Unified Medical System (UMLS) is a terminology integration system developed by the US national library of medicine (NLM) to promote the interoperability of systems and mapping between the multitude of available clinical terminologies [49]. UMLS includes tools for customizing the Metathesaurus (e.g. filter terms by type), for generating lexical variants of concept names (lvg), and for extracting UMLS concepts from text (MetaMap) [49-51]. The Metathesaurus is a large vocabulary database built from electronic versions of various thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research [52]. These are referred to as the "source vocabularies" of the Metathesaurus [52]. It is used in our study to links alternative names and views of the same concept from various ontology systems. The Logical Observation Identifier Names and Codes (LOINC) database provides a universal code system for reporting laboratory and other clinical observations and has been implemented as the standard terminology system for laboratory test / radiology results [53-56]. RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software [57, 58]. Both LOINC and RxNorm are source vocabularies of UMLS and can be linked to SNOMED-CT via UMLS [59-61]. Clinical procedure coding systems including the Current Procedural Terminology (CPT), HCPCS Version of Current Procedural Terminology (HCPT), and Healthcare Common Procedure Coding System (HCPCS) are also source vocabularies of UMLS and can be linked to SNOMED-CT via UMLS [59]. UMLS also contains mapping tables between SNOMED-CT and International Classification of Diseases,

Ninth Revision (ICD-9) / Tenth Revision (ICD-10), which are used for coding diagnosed problems [62].

Many systems have been developed to map heterogeneous terminologies to support communication and semantic interoperability between healthcare centers. STRIDE mapped RxNorm concepts to the SNOMED-CT hierarchy and used the RxNorm relationships in UMLS to link pharmacy data from two EHR sources in the Stanford University Medical Center [60]. Carlo *et al*. classified ICD-9 diagnoses from unstructured discharge summaries using mapping tables in UMLS [63]. Patel and Cimino used the existing linkages in UMLS to predict new potential terminological relationships [64]. Many of these studies only focus on one type of standardized concept. In [6], clinical concepts describing diagnosis, laboratory, and medications were standardized by a concept mapping system which links these concepts to the SNOMED-CT terminologies. These mapped SNOMED-CT concepts were generalized into level 2 nodes for medications and diagnoses and level 3 nodes for laboratories based on the SNOMED-CT concept tree. However, it is much more challenging to standardize clinical procedure concepts and codes due to their comprehensive coverage of products, supplies, services, and surgical and diagnostic procedures.

2.3 UNSUPERVISED LEARNING FOR CLINICAL DATA

Analyzing longitudinal clinical data recorded during care delivery is challenging due to their incompleteness and non-uniformness. Identification of subgroups of patients who experience symptoms with greater or lesser severity [65] or respond to treatment procedures differently may reveal critical risk or treatment factors that impact patient outcome. Laboratory and vitals measurements before, during, and after treatment may act as markers of disease severity [66] and characterize recovery process. Uncovering patient clusters also have prognostic significance – by

constructing cluster-based mortality prediction models, one can achieve superior performance when compared to treating all patient episodes as a single group [67]. However, laboratory and vitals in the form of time series often exhibit different lengths and frequencies due to different syndromes and treatment schedules for different patients. Thus, conventional clustering algorithms aiming to identify patient subgroups cannot be applied directly. Pre-processing methods such as interpolation [68, 69] and resampling [70] can normalize time series data, allowing clustering approaches such as K-means [71] to be used. Alternatively, clustering algorithms have been customized for variable-length time series. These algorithms can be characterized into time-based [72-75], shape-based [76-78], and structural-based [79, 80]. They utilize a variety of similarity (distance) measures such as Dynamic Time Warping (DTW) [78], Longest Common Subsequence (LCSS) [81], Cosine Wavelets [82], Edit distance with Real Penalty (ERP) [83], Global Alignment Kernel (GAK) [84], and Time-Warp Edit Distance (TWED) [76].

While some disease severity can be characterized by a single time series per patient — for example — serum cholesterol levels can be used to characterize conditions of patients with hyperlipidemia [66], others can be better defined by multiple laboratory measurement time series. For instance, systolic blood pressure and diastolic blood pressure should be both considered for patients with hypertension. Clustering approaches for such Multivariate Time Series (MTS) [85] are limited. Existing PCA-based [86], Hidden Markov Model (HMM)-based , partition-based [87], and model-based approaches [88, 89] have been applied to a variety of fields including chemistry and manufacturing, but have not been utilized in clinical settings. This is likely due to the irregularity of clinical time series. As far as we are concerned, clustering approaches have not been developed for MTS with irregular intervals and unequal lengths. These MTS will be referred to as "irregular MTS" throughout this thesis.

2.4 SUPERVISED LEARNING FOR CLINICAL DATA

The use of predictive modeling for clinical decision making has great value in improving outcomes, enhancing patients' experiences, and reducing health care costs [90]. The discriminatory features identified, and the criteria developed via machine learning framework can be used to design and optimize evidence-based treatment plans and to disseminate such knowledge through "rapid learning" across multiple sites [6].

To utilize existing machine learning frameworks, one must first transform clinical variables into suitable forms of input features, or potential predictor variables. Continuous variables including laboratory measurement, vital sign, and drug dosage can be flattened into mean, median, or most-recent-value representations [91-93]. They can also be discretized, clustered, or binarized into "Yes/No" variables [91, 92, 94-96]. Discrete variables including procedure, diagnosis, and prescription can be represented as binary/categorical features [91, 94-96]. In some cases, most recent diagnosis or procedure are isolated as nominal input features [92, 94, 96].

Many state-of-the-art supervised learning algorithms have been utilized in the clinical setting. Among them, Support Vector Machine (SVM) [92, 94, 95, 97-100], Tree-based and Ensemble methods [91, 92, 94, 95, 98, 99], Logistic Regression [91, 92, 95, 96, 100], Naïve Bayes [94, 98], and Artificial Neural Network [97] are the most commonly used in classification tasks. One crucial problem that often rises in the clinical setting is imbalanced datasets, especially when the goal is to maximize the machine learning model's ability to distinguish the minority class [101]. Strategies to address this concern include resampling the dataset [101-103] and modifying the classification algorithm [101].

Feature selection algorithms are essential for narrowing down the list of discriminatory factors in supervised learning models. Reducing the number of predictor variables not only results in faster computations, but also allow for implementation of practical clinical guidelines. The most commonly used methods such as correlation-based feature selection [94, 96], feature importance from tree-based classifiers [91, 92, 97], and Lasso [92, 104, 105] have all shown successes in the clinical setting. Recently, wrapper-based approaches have also gained interests [106, 107]. They use feature selection algorithms to search through the space of attribute subsets with cross-validation accuracy from the classification module as a measure of goodness [107].

2.5 TELEHEALTH AND REMOTE PATIENT MONITORING

Telemedicine is the use of electronic information and communication technologies to provide and support healthcare services to individuals who are some distance from the health care provider [108]. Telemedicine has a growing variety of applications such as telesurgery, psychiatric consultation, and home monitoring of patients [108]. Many health organizations have gained interests in adopting telemedicine technologies to support member physicians' practices or to extend existing services [109]. The use of telemedicine has been rapidly integrated into hospitals [110], home health services [111], and specialty care departments [111, 112].There are over 360 telemedicine programs in the United States and 450 worldwide [113]. These programs most commonly serve those who reside in rural areas, the elderly, and veterans, in many medical specialties [114]. Telemedicine has proven to have many benefits upon implementation. Patients living in rural and isolated regions gain more convenient access to treatment and consultations when they can be conducted online, which eliminates transportation costs and reduce absence of work. Remote patient monitoring achieved through mobile and sensor technologies can improve medical service quality by providing more affordable, accessible, and timely care [115]. The

possible transmission of infectious diseases can also be minimized if patients remain at home in more controllable conditions.

Although many telemedicine intervention programs have been successful, adoption of telemedicine and the level of engagement and services provided across healthcare facilities remain uneven and far from optimal. Compounding the challenges are patient confidentiality and privacy, technical difficulties, licensure issues, and economy and quality of care for reimbursement. There remains an enormous opportunity to expand telemedicine service to provide more timely communication and consultation to patients, reduce the face-to-face demand, and the cost of delivery.

Confidentiality and privacy are major concerns when patients' personal information is exchanged virtually. Breach of personal health information (PHI) can occur on unsecured network, or through illegal access to unencrypted hardware by third parties [116]. This could have potentially destructive consequences such as patient identity disclosure, embarrassment, privacy violation, or integrity violation [117].

Technically challenged staffs and patients create another hurdle to telemedicine practice [116]. Lack of computer literacy and the unwillingness to accept new technology are prevalent among the older generations. In many developing countries, high-speed internet, which is necessary for video conferencing and high-resolution image transmission, is also challenging to implement [118].

Another barrier to telemedicine is that physicians are required to obtain licenses from each state in which their current or potential patients are, or maybe, located [119]. Although some states provide expedite multistate licensing [120], these measures are still insufficient for telemedicine

to expand its impact. It is impractical and expensive for healthcare providers and organizations to obtain medical licenses in all 50 states for telemedicine practice under the current regulations [121]. The Interstate Medical Licensure Compact is an attempt to alleviate the licensing burden of physicians by enacting legislations that promote the expansion of telemedicine without losing control of the regulatory aspects of telemedicine care [122]. It allows the involved states to confront shared needs and issues collectively [122].

Reimbursement issues also present a huge challenge for sustainable telemedicine programs. Medical consultations provided over the telephone are usually not reimbursed by private insurers due to concerns including the abuse of telemedicine services leading to diminished resources in already underserved areas [123]. Only private payers in some states such as Georgia Blue Cross and Blue Shield have guidelines under which telemedical consultations can be reimbursed [123]. In 1997, Medicare started reimbursing for some telemedical services delivered via live interactive video and expanded coverage further in 2000 [124]. Although Medicare reimbursement policy changed little since then, state policies have developed significantly, increasing the number of covered services or payers for telemedicine programs [124]. However, this does not solve the problems for rural areas, which have greater shortages of primary and specialty health care providers, increased prevalence of chronic disease, and a larger population of individuals who rely on Medicare and Medicaid [125].

Since 2009, patient-centered medical homes have been piloted to create a professional care team for coordinated healthcare services to engender a health-friendly environment [5]. Another emerging approach to reducing costs and providing patients with more accessible care is through remote monitoring systems (RMS), which connect patients to providers or health coaches. They can be tailored for the elderly, for the chronically ill, or for patients recently discharged. They can

be used in rural areas where timely and proper healthcare access is limited. RMS allow patients to have more frequent interaction with healthcare providers and expand the reach and influence of healthcare providers to patients between regular office visits. They facilitate proactive patient-centered medical care and advice that is personalized and continuous. They can also be used for tracking patient health and medication compliance. With proper use, there is a potential that RMS can significantly improve quality of services, maintain proactive patient health engagement, while reducing costs. RMS technologies have proven to improve adherence to treatment requirements due to the sense of novelty they provide [126]. Since telehealth has not been as widely adopted and implemented as anticipated due to the challenges listed above, there is a demand for low-cost RMS, especially for low-income patient groups and for the rural population.

CHAPTER III

INFORMATION EXTRACTION AND CONCEPT MINING FROM ELECTRONIC HEALTH RECORDS

3.1 INTRODUCTION

Features of big data revolution can be characterized by the HACE theorem [127]. Big Data is large-volume, heterogeneous, and has autonomous sources with distributed and decentralized control. In addition, there are complex and evolving relationships among data [127]. Big data is the enormous mixture of data in various formats and is thus heterogeneous, which becomes an obstacle when aggregating data in different formats and types [127]. In the clinical context, heterogeneous data result from multi-provider, multi-sites, and multi-terminologies. The autonomous characteristic of big data means that each data source can generate and collect information with no centralized control [127]. Autonomous health information systems have and will continue to play important roles in the clinical setting. With the continuous increase of big data, complexity develops in a high speed [127]. Big data in the EHR is extremely complex due to the scope of information they contain. Finally, big data is always evolving, rendering old-fashion data structure and analytical methodologies insufficient [127]. Clinical big data continues to evolve as it is essential to track the long-term development of patients. Furthermore, medical terminologies are also becoming more and more encompassing and EHR systems are constantly evolving to match the standards.

Kaiser Permanente (KP) uses the Clarity module to transform data from EPIC's operational database into a relational form for reporting. Clarity database from the KP's HealthConnect EHR system stores patient data in over 7,000 tables with over 60,000 columns and updates daily [128]. The EPIC Clarity database has recently been imported to Oracle Exadata for performance

improvement. Structured Query Language (SQL) written in Oracle SQL Developer is the primary programming language used to access the database.

The EPIC Clarity database contains both structured and unstructured data relating to patient demographics, medical history, current and past diagnoses, medications, billings, laboratory orders and results, and clinical notes. Fully extracting the data necessary to understand a cohort of patients is challenging due to the complexity of the database structure and the variety of data entries. In this chapter, we attempt to develop an information extraction framework which is holistic, efficient, accurate, and handles large bodies of unstructured narrative clinical notes.

3.2 MATERIAL AND METHODS

3.2.1 Extract Patient Cohort Characterized by Disease or Symptoms

To extract patient data with certain disease or symptoms, we first utilize the ICD-9 / ICD-10 diagnosis codes. A Patient ID is selected from the problem list table if its corresponding record contains the target diagnosis code(s). In many cases, however, diagnosis codes are not well-maintained, so it is necessary to utilize billing information, laboratory measurement data or narratives in clinical notes for more accurate case detection. This can be done using semantic matching of key terms describing medical conditions. The extracted patient IDs are then used to link to the other data tables to extract the relevant information.

Table 3.1 lists the types and coverages of information extracted. Although most demographics, medications, billings, procedures, and co-existing conditions can be found directly from structured data tables, encounter-level data containing physician visits and referrals, dietary management, and suspected problems must be extracted from the clinical notes table.

TABLE 3.1 DATA COVERAGE BY SOURCE TABLES.

Coverage	Source database tables
Encounter-level data	Encounter / Clinical notes tables
Medications data	Medications table
Billing information	Billings table
Procedures	Billings / Clinical notes tables
Clinical notes	Clinical notes table
Problem list (co-existing conditions)	Problem list / Billings/ Clinical notes tables
Laboratory	Order table / Clinical notes table

3.2.2 Extract Patient Cohort Characterized by Treatment Features

To extract patient data with certain treatment features (i.e. procedures, prescriptions, laboratory measurements), all the possible vocabularies that represent the treatment features must first be identified. These vocabularies are compiled into a list and are used to index the billings / laboratory / medications tables to select the target patient IDs. Alternatively, regular expressions can be used to represent groups of vocabularies to create more succinct queries.

3.2.3 Table Partitioning and Temporary Views

In many data extraction tasks, the targeted patient cohort contains millions of patient records amounting to terabytes of data. In such cases, table partitions are created to retrieve data by chunks and reduce local storage loads. Temporary views are used to reduce server loads.

3.2.4 PHI Encryption for Structured Data and Narrative Texts

The SHA-256 Cryptographic Hash Algorithm is used to encrypt Patient IDs contained in every data record. For unstructured free-text data, the transition-based parsing model implemented in the Python spaCy package [129] is applied to detect and de-identify PHI in clinical notes. Entities including “PERSON”, “NORP”, “ORG”, and “GPE” are identified. These entities cover patient names, nationalities, organizations, and addresses. In addition, a regular expression-based filter is leveraged to replace telephone numbers.

3.2.5 Information Extraction from Narrative Clinical Texts

An end-to-end “pipeline” (software from EPIC Systems coded to process data into a more usable form) is developed for extracting key clinical features from narrative documents. These features are then filtered by negation detection and remaining features are mapped to standardized SNOMED-CT terminology. Figure 3.1 shows the feature extraction pipeline from clinical text. The content summarization module is implemented based on the TextRank algorithm [130]. The CLiNER concept recognition model [26] is leveraged to extract key clinical features including problems, procedures, and tests. An improved Negex [29] algorithm is then used to filter features within a negated context. This portion of the pipeline has been implemented as a secure in-house web application where the user is allowed to upload files containing narrative clinical notes and obtain extracted key clinical features along with their contexts.

From here, two approaches can be taken: 1) utilize MetaMap to map the extracted features to the SNOMED-CT terminology system and filter out features that are not mapped. The hierarchical structure of SNOMED-CT and MetaMap are utilized to remove general concepts (e.g. “Body structure”, “Clinical finding”, “Biological agent”) that are situated at the top two levels in the SNOMED-CT concept tree; 2) utilize the terminology mapping system developed in chapter 4 to directly map these concepts to SNOMED-CT. These standardized concepts can be consolidated into input features that could be directly input into machine learning algorithms for knowledge discovery.

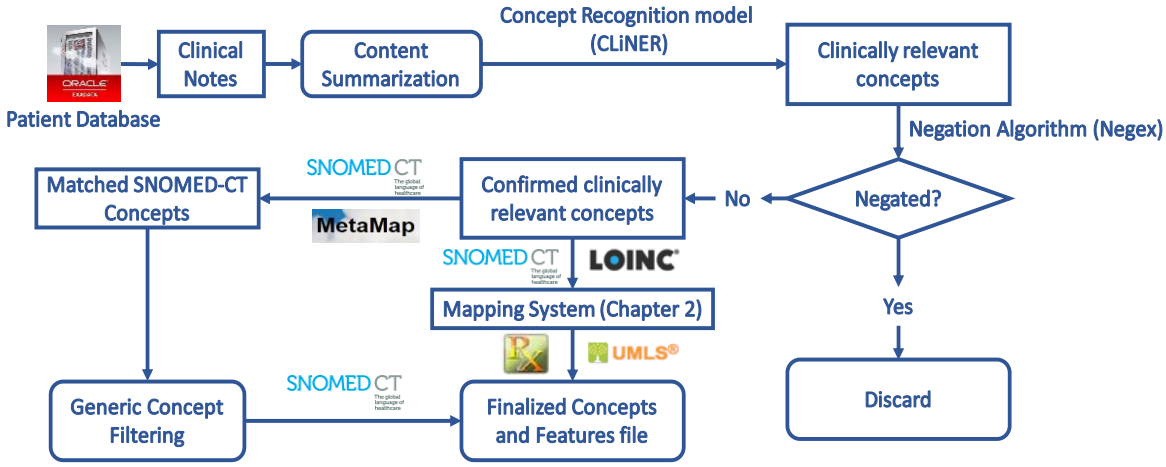


Figure 3.1 Treatment feature extraction from clinical texts.

3.3 RESULTS

In this section, we demonstrate the use of the EHR information extraction methods for two different patient cohorts: prostate cancer and chronic kidney diseases (CKD).

3.3.1 Patients with Prostate Cancer

Prostate cancer is the most frequently diagnosed cancer in 105 countries and the fifth leading cause of cancer death in men [131]. It is estimated that there will be 174,650 new cases of prostate cancer in the US in 2019 and an associated 31,620 deaths [132]. Early prostate cancer detection has been achieved through prostate-specific antigen (PSA) testing and biopsy of tissue removed during prostatectomy or at autopsy [131]. However, PSA testing has been questioned for its accuracy in detecting prostate cancer [133] and for the high cost it carries in terms of overdiagnosis and overtreatment [134-137]. Through mathematical modelling, [138] concluded that under the assumption that stage shift implies survival shift–which motivates early detection of cancer, PSA screening likely explains half or more of the mortality reduction observed in the US since the early 1990s.

EHR provides long-term tracking of patient PSA testing results. These longitudinal data can be extracted given the lab component IDs or names corresponding to the testing procedure. The rate of increase in PSA level, often represented using PSA doubling time or PSA velocity, has been widely used in the management of prostate cancer [139]. Trajectories of post-treatment PSA is associated with cancer mortality [140]. In this thesis, a clustering technique is developed and applied to identify patient subgroups with distinct post-treatment PSA trajectories.

The extracted dataset from EPIC EHR covers 98,806 patients with the ICD-9 code 790.93 or the ICD-10 code 97.20, "elevated prostate specific antigen (PSA)". This dataset spans the years 1997-2018 and is composed of patient-level data (70Mb), problem lists (384Mb), medications (7.3Gb), billings (167Mb), laboratory orders (10Gb), and clinical notes (46.1Gb), totaling 64.02 Gigabytes. Patient IDs were successfully encrypted using SHA-256 encryption. PHI including patient names, addresses, institutions, age, phone numbers, and email addresses were detected and encrypted into dummy tokens.

The clinical concept extraction system was applied on a subset of patients treated with radioactive seed implants. An additional 2,194 standardized clinical features were extracted from their clinical notes, including "Chronic pain syndrome", "Placement of stent", "Nerve conduction testing", "Vascular Calcification", "Overweight", "Obstructive sleep apnea syndrome", "Neoplasm, metastatic", and "Lithotripsy", etc.

Patient PSA laboratory test results were used as indicators of disease severity. PSA records were retrieved by the following method: 1) component IDs for lab records matching the query string "%PSA%" were retrieved; 2) PSA-irrelevant lab components were discarded, leaving 10 unique component IDs corresponding to "PSA-screening", "PSA-monitoring", "PSA", "PSA

FREE", "PSA % FREE", "PSA, external result", "PSA, MHS", "PSA with reflex FPSA, external result", "PSA, screening", and "PSA, cancer monitoring"; 3) "PSA FREE" and "PSA % FREE" were removed from the list of candidate components since free PSA is reported as a percentage of the total that is not protein bound, i.e., free. The higher the free PSA, the lower the likelihood of cancer; 4) PSA lab records were then retrieved by patient IDs and the filtered component IDs; 5) Missing, erroneous, and duplicated records were removed, and the remaining records were sorted by date and transformed into time series format for each patient.

3.3.2 Patients with Chronic Kidney Disease (CKD)

Kidney is an important organ of human body – filtering blood, removing waste, balancing fluid, and controlling the level of electrolytes. Chronic Kidney Disease (CKD) is becoming more prevalent at a rapid speed across the world.

CKD can be divided into 5 stages based on estimated glomerular filtration rate (eGFR) measurement. Early diagnosis of CKD prevents patients from regressing into late-stage CKD which causes serious complications. Late-stage CKD can lead to end-stage renal disease (ESRD) and cardiovascular disease (CVD), which steeply increase patient pain and economic burden. However, the gradual loss of kidney function is difficult to diagnose due to the absence of direct evidence from clinical trials [141]. Hence frequent and regular measurements of serum creatinine—used to calculate eGFR—is essential for evaluating changes in renal functions. Identifying trends in eGFR is more important than one-off readings, as suggested by the Renal Association, "a progressive fall in eGFR across serial measurements is more concerning than stable readings which don't change over time" [142]

EHR provides a possibility for health care organization to identify early-stage CKD. Lenart *et al*. developed clustering techniques to detect progression of CKD [143]. K-medoids clustering was applied on patients' routine measurements and lab tests such as blood pressure, body mass index, Hemoglobin A1c (HbAlc), triglycerides and high-density lipid cholesterol [143]. The Cluster Progression Score (CPS) was designed to measure the patients' relative health status [143]. This clustering process can help health care organization diagnose early stage CKD by monitoring the recorded lab measurements.

The extracted dataset from EPIC EHR covers 33,303 patients with the ICD-9 code starting with "585" or ICD-10 code starting with "N18", both referring to "Chronic Kidney Disease". This dataset spans the years 1997-2018 and is composed of patient-level data (24Mb), problem lists (288Mb), medications (6.74Gb), billings (1.90Gb), laboratory orders (8.66Gb), and clinical notes (18.55 Gb), totaling 36.16 Gigabytes. Patient IDs were successfully encrypted using SHA-256 encryption. PHI including patient names, addresses, institutions, age, phone numbers, and email addresses were detected and encrypted into dummy tokens.

Patient eGFR laboratory test results were used as indications of disease progression. eGFR records were retrieved by the following method: 1) component IDs for lab records matching the query string "%eGFR%" or "%GLOMERULAR FILTRATION RATE%" were retrieved; 2) Irrelevant lab components were discarded, leaving 16 unique component IDs. eGFR records matching these component IDs are examined it was discovered that only records corresponding to two component IDs "12122727" and "12122728" were well-maintained in the EHR. 3) eGFR lab records were then retrieved by patient IDs and these two component IDs. 4) Missing, erroneous, and duplicated records were removed, and the remaining records were sorted by date and transformed into time series format for each patient.

3.4 DISCUSSIONS

In this chapter, a comprehensive information extraction pipeline for EPIC-based EHR system is designed. This pipeline is capable of effectively and efficiently extracting structured data from interlinked data tables. It can also identify key clinical concepts including problems, procedures, and tests along with their word sense. These concepts can be further standardized and generalized using the SNOMED-CT terminology system. This pipeline is applied to two cohorts of patients – those with prostate cancer and chronic kidney diseases, to generate tabularized data files with standardized terminologies and reduced feature dimensions. These data files can be directly provided as input into machine learning algorithms for further knowledge discovery.

With minor modifications to some processes in the pipeline, it can be applied to similar large EHR datasets and for other patient cohorts. These modifications might include: 1) redesigning SQL queries by modifying diagnosis codes when extracting patient ID list to accommodate different target cohorts; 2) modifying SQL queries to extract additional data from target disease-specific tables; 3) reidentifying new motifs through expert recommendation and/or manual exploration of free-text data and redesigning new regular expressions for pattern-based feature extraction.

Through the design and implementation of this pipeline, many big data challenges including volume, variety, veracity, and value are addressed. This results in a highly robust, efficient, and customizable pipeline that can be easily applied to current EHR databases to fulfil their potential in both academic and clinical research.

CHAPTER IV

ESTABLISH DATA INTEROPERABILITY WITH MEDICAL TERMINOLOGY MAPPING

4.1 INTRODUCTION

EHR data are recorded by different clinicians and different providers at various hospital sites. As a result, data heterogeneity becomes a major issue due to the significant practice variation in style of reporting, use of terminologies and descriptive content. These pose a huge challenge for understanding these already complicated data. Tackling the problem of data heterogeneity in the EHR is essential for conducting predictive analytics and discover knowledge.

To the best of our knowledge, integration of EHR data across hundreds of healthcare sites and millions of patients has not been attempted previously. Past studies have focused on inter-terminology mapping between specific source vocabularies. In this chapter, a terminology mapping system is developed for an EHR covering 737 clinical sites and de-identified data for over 2.7 million patients with data collected from January 1990 to December 2012. This system establishes interoperability among these multi-site data by accurately mapping ICD-9 diagnosis codes and free text describing laboratory and medication terms to concise structured SNOMED-CT medical concepts. This allows for shared characterization and hierarchical comparison of patient characteristics. This system is then extended to include the functionality of mapping CPT/HCPCS codes and free text describing clinical procedures to SNOMED-CT concepts, and an automated pipeline is developed to streamline the mapping process to allow easy adaptation to other EHR systems.

4.2 MATERIAL AND METHODS

4.2.1 Dataset Description and Management

This study utilizes EHR data of 2.7 million patients collected from 737 healthcare facilities. A relational database is first designed with Postgres 9.2.18 to store these data. Thirteen tables are created containing patient records pertaining to procedures, demographics, diagnosis codes, laboratory measurements and medications. Indexes are developed for each table to enable rapid search and table joins for data querying. The data size for indexes is an additional 11 GB, totaling 27 GB for the entire database. It is labeled as the CCI-health database, where CCI stands for Care Coordination Institute. In the CCI-health database, 2.46 million patients are associated with a diagnosis, 1.89 million are associated with procedures, 1.33 million are associated with laboratories, and 955,000 are linked to medications [6].

4.2.2. Data Integration and Mapping to Standardized Medical Concept

Laboratory and medication records are described with free text entries without unique codes for each entry. Since clinicians may describe identical treatments with many possible variations, it is essential to map entries to structured concepts without ambiguity. Overall 803 unique lab phrases and 9,755 medication phrases were extracted from the patient records. In this study, Metamap is used to map laboratory and medication terms to UMLS Metathesaurus terms. Laboratory and medication terms are then linked to LOINC and RxNorm terms respectively. This is done respectively using the UMLS MRCONSO and RXNCONSO tables. The final step is to map associated terms to concepts in the SNOMED-CT ontology. LOINC and RxNorm terms established from the CCI-health database are linked to SNOMED using the UMLS MRREL and RXNREL tables. In this implementation, for LOINC, only concepts that have the name “procedure” were returned from the MRREL table. For RxNorm, only concepts that have “has_form”, “has_ingredient”, and “has_tradename” relationships are returned from the RXNREL table. When

medication entries in an EHR and a SNOMED concept are named completely differently, relationships can still be found due to rules such as tradenames and ingredients. Figure 4.1 and Figure 4.2 show the workflows for mapping laboratory and medication phrases to SNOMED-CT concepts.

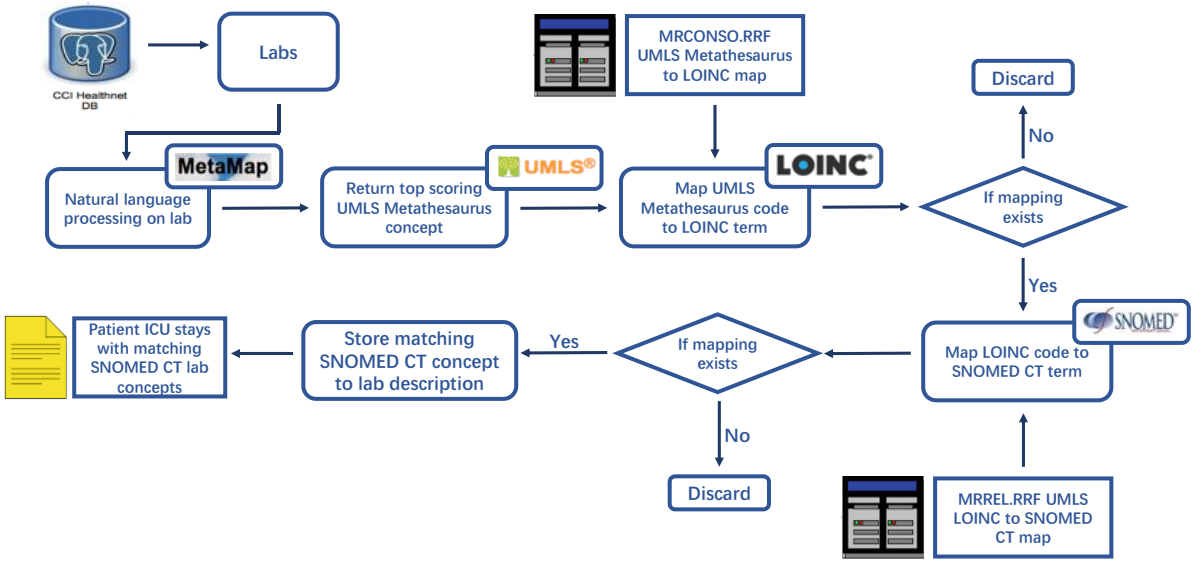


Figure 4.1. Mapping process for laboratory phrases to SNOMED-CT concepts.

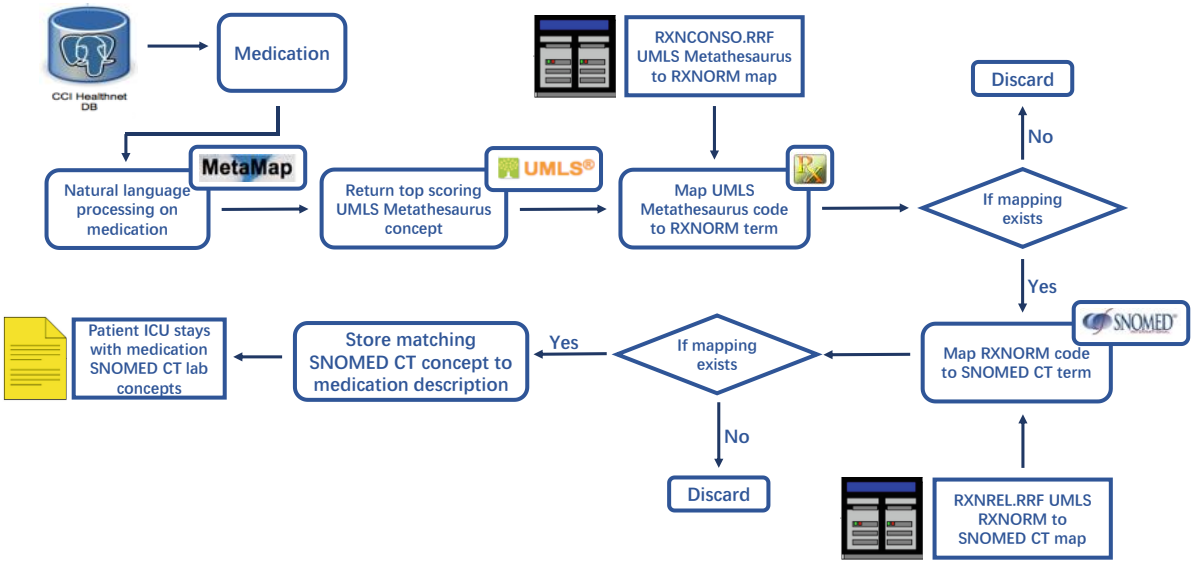


Figure 4.2. Mapping process for medication phrases to SNOMED-CT concepts.

The CCI-health database employ ICD-9 codes for patient diagnoses [53]. This makes the mapping procedure to SNOMED-CT concepts slightly different from those designed for laboratories and medications. The ICD9CM_SNOMED_MAP table in UMLS can be used to map ICD-9 directly to SNOMED-CT concepts. However, this does not include all ICD-9 codes that are associated with patients in the CCI-health database. Metamap is then used to analyze the descriptions of the remaining ICD codes that are not found in the ICD9CM_SNOMED_MAP table and map them to UMLS Metathesaurus concepts. The MRCONSO table is then used to map the UMLS concepts to associated SNOMED-CT concepts.

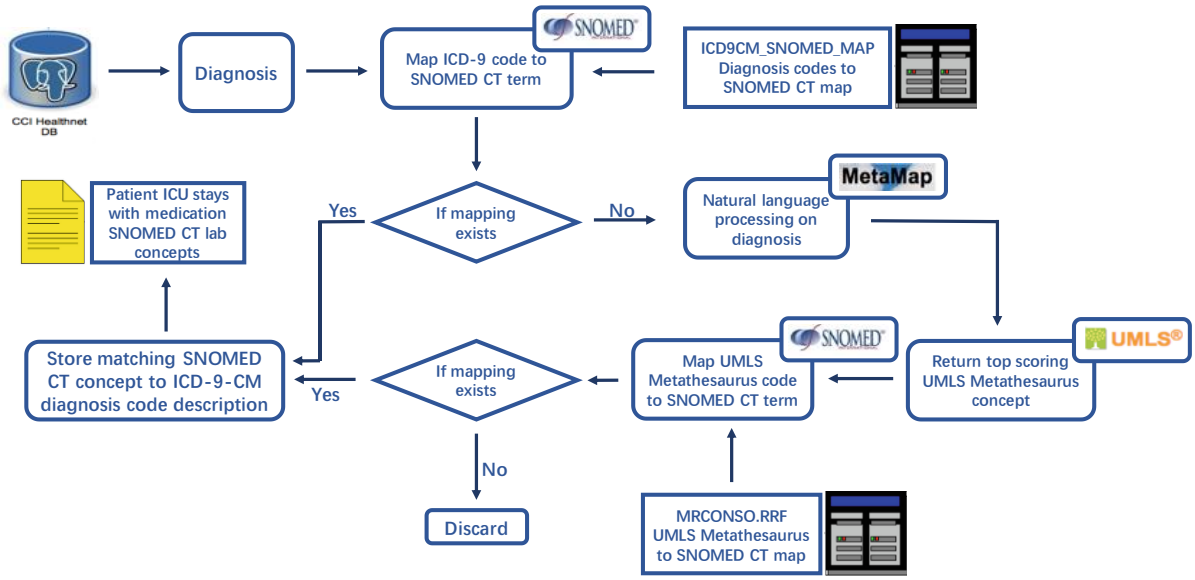


Figure 4.3. Mapping process for diagnosis ICD-9 codes to SNOMED-CT concepts.

Procedures are recorded in the EHR database using the Healthcare Common Procedure Coding System (HCPCS), which consists of two levels including the level I common Current Procedural Terminology (CPT) codes and Level II codes which identify products, supplies, and services not included in CPT [144]. Mapping procedures to SNOMED-CT requires a more holistic approach than that of the other concepts since HCPCS contains codes for clinical labs, medications,

and procedures/interventions [145]. Although HCPCS is one of the source vocabularies in the UMLS, many codes do not have direct matches to SNOMED-CT concepts. Therefore, the following mapping process is developed: first, concepts strings corresponding to the procedure codes are extracted from the MRCONSO table. Out of the 11,374 unique codes, 7,570 had one or more matches in the MRCONSO table. Secondly, MetaMap is used to identify one or more UMLS candidate concepts that are associated with these extracted concepts. Candidate concepts that had the highest evaluation score using MetaMap's natural language processing are selected. Some of these UMLS concepts have corresponding Metathesaurus concepts that could be linked to SNOMED-CT concepts directly. For those UMLS concepts that do not have a direct mapping to SNOMED-CT, LOINC and RxNorm terms are used to link them to SNOMED-CT using MRREL and RXNREL tables. If a UMLS concept can be linked to both LOINC and RxNorm sources, the source with the higher matching score is selected.

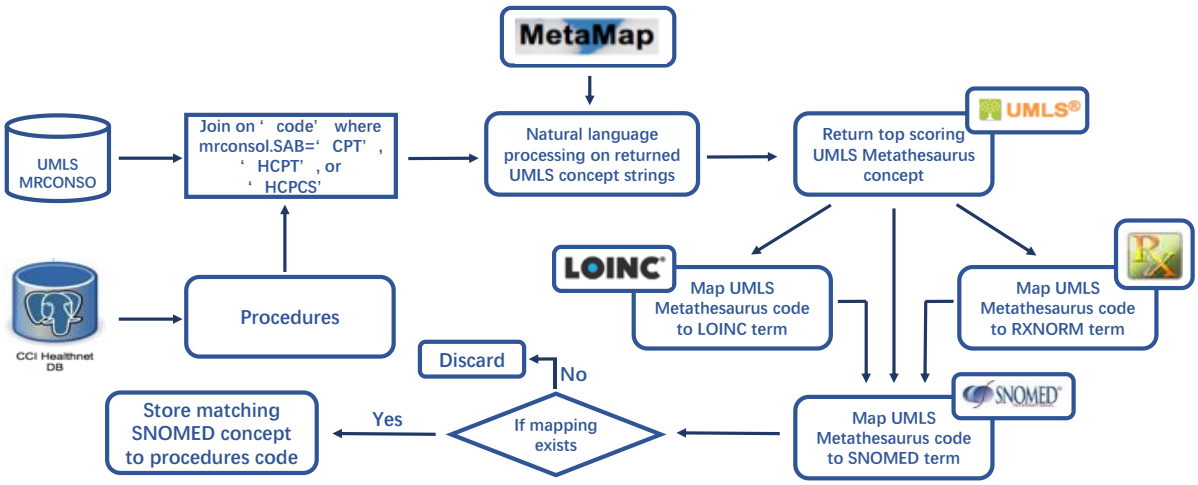


Figure 4.4. Mapping process for procedure codes to SNOMED-CT concepts.

4.2.3 Automation of Mapping Process

Although the mapping approaches for labs, medications, diagnoses, and procedures are different, they share similar intermediate steps and can be automated. In this section detailed instructions are

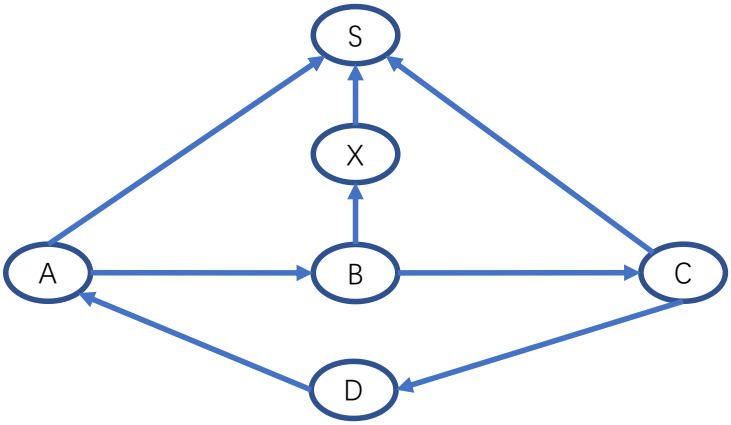
provided on how to set up the necessary knowledge databases and software, and how the developed scripts are leveraged to establish the mapping tables.

The Postgres 9.2.18 database is set up on a Red Hat Enterprise Linux Server release 7.3 operating system. MetamorphoSys, the UMLS installation tool, is downloaded from the NLM website (2016AB release) and used to generate the UMLS data in the Rich Release Format (RRF) files. Database table creations scripts can be found inside the installation directory. The Postgres table and index creation scripts provided by UMLS are used to load the RRF files into the database. Metamap 2016 v2 binaries are downloaded from https://metamap.nlm.nih.gov/ and installed. After setting up the EHR database, the UMLS knowledge databases, and Metamap, a script is developed in Python 2.7.5 to streamline the mapping procedure. It is readily deployable after specifying the input file containing terms to be mapped, UMLS database access credentials, the path to Metamap binaries, and the mapping output database and output file names. The script will read, analyze, and store matching SNOMED-CT concepts to labs, medications, diagnoses, and procedures in the EHR for rapid retrieval and further expert evaluation.

4.2.4 Refinement and Validation of Mapping Results

SNOMED-CT provides a rich hierarchy enabling semantic distances between concepts to be measured by path distances. ANeo4j Graph Database is developed for the CCI-health data to rapidly compute common ancestor queries between the mapped SNOMED-CT terms. In our Neo4j Graph Database, tab-delimited files of all SNOMED concepts and relationships are exported from SNOMED CT Postgres relational database tables. The tab delimited files were then directly loaded into Neo4j community edition 2.0.0 using their batch inserter (http://docs.neo4j.org/chunked/milestone/batchinsert.html). This results in a SNOMED Graph Database that has many cycles. The cycles greatly impede graph operations such as returning all

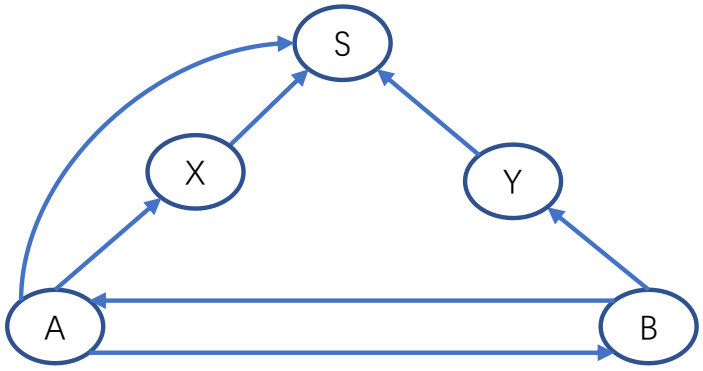
paths between nodes. This issue is addressed by removing edges from these cycles to construct a directed acyclic graph (DAG). To select the proper edge that needs to be removed from each cycle, three different criteria are used: 1) maximum depth, 2) shortest path, and 3) fan-in count. Maximum depth is the maximum number of nodes between the current node and head node (root) of the ontological graph. Here the node with the Concept ID "138875005", corresponding to the Concept Name "SNOMED CT Concept", is chosen as the head node. In each cycle, edges whose source node has a smaller maximum depth than its target node is removed (Figure 4.5).



Nodes in Cycle	Maximum Depth
A	3
B	4
C	5
D	4

Figure 4.5. Maximum Depth Criteria: S is the head node. In the cycle A->B->C->D->A, since A has a smaller maximum depth than B, and B has a smaller maximum depth than C, edges A->B and B->C will be removed.

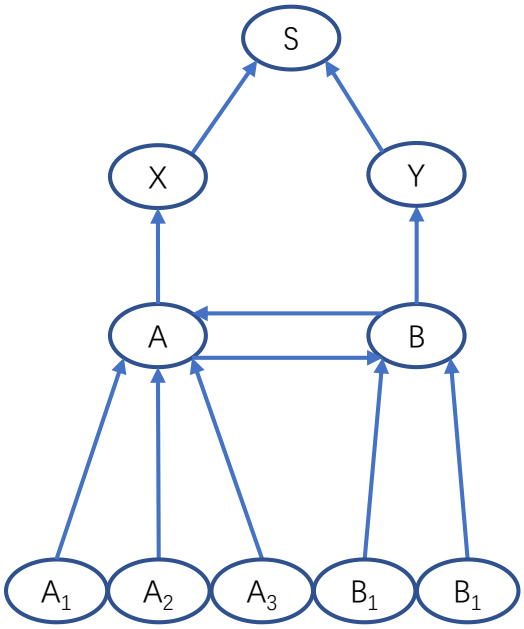
Shortest path is the shortest distance (number of edges) from the current node to the head node. In cases where the maximum depth of all nodes in the cycles are equal, edges whose source node has a shorter shortest path than its target node is removed (Figure 4.6).



Nodes in Cycle	Shortest Path
A	1
B	2

Figure 4.6. Shortest Path Criteria: S is the head node. In the cycle A->B->A, A and B has equal maximum depth. Since A has a shorter shortest path to S than B has, edge A->B will be removed.

Finally, if all nodes in the cycle have equal maximum depths and shortest paths, the Fan-in count criteria is used. Fan-in count is the number of incoming edges to the current node in the cycle. Nodes with larger degrees of incoming nodes are expected to represent more generalized concepts. Therefore, edges whose source node has a larger fan-in count than its target node are removed (Figure 4.7).



Nodes in Cycle	Fan-in Count
A	4
B	3

Figure 4.7. Fan-in Count Criteria: S is the head node. In the cycle A->B->A, A and B has equal maximum depth and shortest path. Since A has a larger fan-in count, edge A->B will be removed.

These criteria are selected to reduce the likelihood of a node representing a more specific concept from becoming a parent of a node representing a more generalized concept in the final concept graph.

After an acyclic SNOMED-CT Graph database is created, graph computations such as shortest paths and common ancestor queries can be performed rapidly. This is beneficial since laboratories, diagnoses, and medications are all mapped to many SNOMED-CT concepts that can

be too specific for machine learning analysis. In this study, all nodes are assigned a depth level according to the minimum number of edges that must be traversed to reach the root node. The root node is at depth level 0. All the mapped SNOMED-CT concepts can then be generalized to concepts at a higher depth level [6]. It is important to choose an appropriate depth level to accurately distinguish patient characteristics from one another. For medications and diagnosis, a depth level of 2 is chosen. A depth level of 3 is chosen for laboratories, since assigning lower depth levels returned concepts that are too general [6]. On the other hand, because procedure terms can describe a whole spectrum of clinical terminologies and are mapped to SNOMED-CT via multiple source vocabularies, concept generalization is not performed, and mapping results are reserved for expert evaluation to ensure coverage and accuracy. For a given SNOMED-CT concept, Neo4j can quickly calculate all possible paths to the root node. With the Cypher query language, Neo4j returns all nodes for a given depth level that are crossed from all possible paths to the root of the hierarchy. After converting all SNOMED-CT concepts mapped from laboratory, medication, and diagnosis, concepts that are too general and do not provide key information in characterizing patients are removed. Examples include “Disorder of body system (362965005)”, “Measurement procedure (122869004)”, “Chemical procedure (83762000)”, and “Types of drugs (278471000)”.

4.2.5 Running Time Analysis

A running time analysis of the automated mapping pipeline is performed. Since the pipeline is composed of multiple steps utilizing multiple algorithms, the running time varies with input parameters. However, it can be shown that the entire process scales linearly with the number of input vocabularies. In addition, we empirically analyze the scaling behavior of the pipeline by conducting several computational experiments.

The most time-consuming step of the mapping pipeline is Metamap's natural language processing step during which the input text undergoes a lexical/syntactic analysis consisting of 1) tokenization, 2) part-of-speech (POS) tagging, 3) lexical lookup of words in SPECIALIST lexicon, and 4) identification of phrases and lexical heads by the SPECIALIST parser (26). Each phrase found by this analysis is further analyzed by 5) variant generation, 6) candidate identification, and 7) mapping construction (26).

Let n be the number of input vocabularies / phrases. The asymptotic computational complexity of each step of the analysis is given below:

1) The tokenization module uses a regular expression-based approach which needs O(n) time (27);

2) POS tagging is implemented with the Viterbi Algorithm (28), which uses a bigram Hidden Markov Model with 60 possible states corresponding to the POS tags (29). The asymptotic complexity of the Viterbi Algorithm is $O(n|V|^2)$ (28), where V=60 — the number of possible POS tags;

3) The lexical lookup is a simple dictionary lookup process which takes O(n) time;

4) Identification of phrases and lexical heads also takes O(n) time;

5) Variant generation takes O(mn) time, where m is the number of words in the lexicon (27);

6) Candidate retrieval is dependent of the number of variants generated in the previous step. This step consists of a) tokenization of the variants and b) lexical lookup for the first canonical-

form token. Both steps use O(k) time, where k is the number of variants generated. Therefore, candidate retrieval takes a total of O(nk) time;

7) The mapping construction step evaluates each candidate and generates a Matchmap (27). This step uses O(cn) time, where c is the number of candidates retrieved in the previous step.

Since V, m, k, and c can all be considered as constants, and each step above scales linearly with the number of input vocabularies n, the total running time of Metamap processing uses linear time O(n). Remaining steps of the mapping pipeline involves selecting the concept type filtering, source vocabulary matching, and filtering, each of which also uses linear time O(n). Thus, the entire mapping pipeline has an asymptotic runtime complexity of O(n).

To demonstrate the efficiency of the mapping pipeline, computational experiments are conducted as follows: the pipeline was implemented and executed using Python 3.6.8 on a Red Hat Enterprise Linux 7.3 (Maipo) Server with 8 GenuineIntel Intel Xeon E3-12xx v2 (Ivy Bridge, IBRS) processors. 15,000 unique phrases were extracted from the database, with 5,000 related to each of medication, laboratory, and procedure (corresponding to CPT code). The mapping pipeline was executed for each category of phrases using incrementing subsets of 1000, 2000, 3000, 4000, and 5000 and the CPU times are recorded for each subset. Since mapping for diagnosis codes can be achieved without the Metamap processing step, they are not included in the experiments. Table 4.1 compares the CPU time of each set of experiments. It is evident that the processing time increases linearly with the number of input phrases within each category with notable overhead. Mapping of each input phrase takes on average between 1-1.6 seconds. Thus, the mapping pipeline is highly efficient and scales linearly with the number of input phrases.

TABLE 4.1. RUNTIME OF MAPPING PIPELINE FOR MEDICATION, LABORATORY, AND PROCEDURE PHRASES IN CPU-SECONDS.

	CPU Time (seconds)		
Number of Input Phrases	**Medication**	**Laboratory**	**Procedure**
1000	1101.10	1297.99	1285.66
2000	2033.56	2356.91	2227.67
3000	3028.25	3346.80	3329.89
4000	4034.20	4660.72	4375.34
5000	5000.90	5626.45	5349.73

4.3 RESULTS

Free text entries from laboratories, medications and ICD9 codes are all successfully mapped to SNOMED-CT concepts. Of the 803 unique lab phrases in the CCI-health database, 603 can be linked to SNOMED-CT; this covers 1.20 million of 1.33 million patients. Similar successes are found for concept mapping of medications, procedures, and diagnoses: 5,899 of 9,755 medication phrases are mapped to SNOMED-CT; this covers 801,025 of 952,729 patients. 7,609 of 11,374 procedure codes are mapped to SNOMED-CT; this covers 1,793,973 of the 1,894,333 patients. 10,655 of 29,371 ICD-9 codes are linked to SNOMED-CT concepts. Specifically, 2.35 million of 2.46 million patients contain at least one ICD-9 diagnosis code that can be mapped to SNOMED-CT. Table 4.2 shows examples of medical entries for medications, laboratories, diagnoses, and procedures and their mapped SNOMED concepts. Medications can have many brand names and a range of ingredients. Laboratory and procedures may also be described by physicians with a great deal of variations. Table 4.3 shows how variations and ambiguity among phrases are eliminated by our mapping algorithm.

TABLE 4.2. MEDICATION, LABORATORY, PROCEDURE, AND DIAGNOSIS PHRASES TO SNOMED CONCEPT MAPPING.

Entity type	Entry	SNOMED concept	SNOMED code
Medication	Mucinex D Tablets Extended Release	Pseudoephedrine (product)	91435002
Medication	Bromfed Dm Cough Syrup		

TABLE 4.2 continued

Medication	Lortab Tablets	Acetaminophen (product)	90332006
Medication	Roxicet Tablets		
Medication	Hydrocortisone Acetate	Hydrocortisone preparation (product)	16602005
Medication	Proctofoam Hc Aerosol		
Lab	lipid-triglycerides	Plasma lipid measurement (procedure)	314039006
Lab	lipid-vldl-cholesterol		
Lab	urine-creatinine-24hr	Urine creatinine measurement (procedure)	271260009
Lab	urine-creatinine-random		
Lab	glucose tolerance test (3 hr)	Glucose tolerance test (procedure)	113076002
Lab	glucose-tolerance-test-5-hour		
Procedure	Removal of abdominal lining, uterus, both ovaries and fallopian tubes, and pelvic and aortic lymph nodes with tumor reduction	Abdomen incision (procedure)	108188006
	Removal of prosthetic material or mesh, abdominal wall for infection (eg, for chronic or recurrent mesh infection or necrotizing soft tissue infection) (List separately in addition to code for primary procedure)		
Procedure	Re-exploration of liver wound with removal of packing	Exploration of liver (procedure)	265437009
	Exploration of hepatic wound with coagulation and packing of liver		
Procedure	Fecal blood scrn immunoassay	Immunoassay method (procedure)	414464004
	Analysis of substance using immunoassay technique		
Diagnosis	V13.64 Personal history of (corrected) congenital malformations of eye, ear, face and neck	Congenital deformity (disorder)	276655000
Diagnosis	V13.65 Personal history of (corrected) congenital malformations of heart and circulatory system		
Diagnosis	770.86 Aspiration of postnatal stomach contents with respiratory symptoms	Respiratory symptom (finding)	161920001
Diagnosis	770.14 Aspiration of clear amniotic fluid with respiratory symptoms		
Diagnosis	806.35 Open fracture of T7-T12 level with unspecified spinal cord injury	Open fracture (disorder)	397181002
Diagnosis	806.11 Open fracture of C1-C4 level with complete lesion of cord		

TABLE 4.3. MEDICATION, LABORATORY AND DIAGNOSIS SNOMED DEPTH LEVEL MAPPING.

Entity type	SNOMED concept	SNOMED code	SNOMED generalized concept	SNOMED general code	Count
Medication	Zolpidem (product)	96231005	Psychotherapeutic agent (product)	46063005	104521
Medication	Risperidone (product)	108386000			52970
Medication	Azithromycin (product)	96034006	Antibacterial drugs (product)	346325008	66150
Medication	Amoxicillin (product)	27658006			64842
Medication	Valsartan (product)	108581009	Hypotensive agent (product)	1182007	117287
Medication	Benazepril (product)	108572003			98706
Lab	Urine screening for protein (procedure)	171247004	Evaluation of urine specimen (procedure)	442564008	4267836

TABLE 4.3 continued

Lab	Measurement of fasting glucose in urine specimen using dipstick (procedure)	442033004			2799670
Lab	Hemoglobin variant test (procedure)	302763003	Hematology procedure (procedure)	33468001	18262531
Lab	Red blood cell count (procedure)	14089001			16487444
Lab	Histamine release from basophils measurement (procedure)	63987006	Immunologic procedure (procedure)	108267006	145032
Lab	C-reactive protein measurement (procedure)	55235003			79804
Diagnosis	Tongue tie (disorder)	67787004	Congenital anomaly of gastrointestinal tract (disorder)	128347007	1717
Diagnosis	Idiopathic congenital megacolon (disorder)	268209004			1177
Diagnosis	Hemorrhage of rectum and anus (disorder)	266464001	Hemorrhage of abdominal cavity structure (disorder) Hemorrhage of abdominal cavity structure (disorder)	444312002	41580
Diagnosis	Hematoma AND contusion of liver without open wound into abdominal cavity (disorder)	24179004			1031
Diagnosis	Acute bronchiolitis due to respiratory syncytial virus (disorder)	195739001	Disorder of the larynx (disorder)	60600009	9710
Diagnosis	Vocal cord palsy (disorder)	302912005			5479

4.4 DISCUSSION

In this chapter, interoperability among EHRs from 737 clinical facilities is established using a mapping process that disambiguates free text entries. The process provides a unique way to link to structured medical concepts despite the extreme variations that can occur during clinical diagnosis and documentation. It enables more powerful systems to be developed for future studies where semantic distances can be calculated between patient records due to their association with hierarchical concepts [6]. The graph database allows for rapid data access and queries. The automation of the mapping pipeline allows for easy adaptation to other EHR systems after providing minimal user input. Most importantly, this mapping system can be effectively applied in clinical big data analytics. It improves the efficiency and discriminatory power of existing

machine learning algorithms and allows them to extract useful knowledge from clinical data that facilitate the design and implementation of optimized evidenced based treatment plans. Data interoperability established via the mapping system can also facilitate the design of universal clinical guidelines and rapid knowledge dissemination.

CHAPTER V

CHARACTERIZING PATIENT TREATMENT OUTCOMES BASED ON LONGITUDINAL DATA

5.1 INTRODUCTION

Pattern recognition from longitudinal time series data is important in many domains. In the clinical setting, time series data are generated during the delivery of care and exhibit characteristics such as irregular sampling rate and unequal lengths. These include laboratory measurements, vital signs, and prescription dosages, and are extremely valuable because they track the long-term progression of the development of disease. The scale and length of time-series varies a lot because different patients have different syndromes [146]. One technique for pattern recognition is transforming continuous time-series data into discrete form and using feature extraction. Zhao *et al*. uses each clinical events' temporal weight to create features, which can be used as classifiers [147]. In [146], Zhao *et al*. applied machine learning algorithms to transform longitudinal time-stamped clinical data to tabular format and a subsequence-based method to detect adverse drug events using clinical measurements. In another study, Yuan *et al*. proposed an end-to-end model, Wave2Vec, which combines deep learning and NLP to learn inherent and temporal signals without requiring any expert medical knowledge [148]. The model was applied to two clinical datasets to improve interpretability of time series data. These studies demonstrate that discrete features can be extracted from time series and used to uncover patterns that lie within these longitudinal data.

Clustering analysis is another effective approach for pattern detection from time series data. Three main categories of approaches are used to cluster time series data [149]. The first approach is to convert time series into simple objects and apply conventional clustering algorithms on these objects. For example, one possible method is to shrink the time series to a single point, which is

assigned with some analytical value of raw data points [150]. The second approach is to customize existing clustering algorithms so that they are applicable to the unconventional time series data. Usually, distance metrics for these clustering algorithms need to be modified. For example, dynamic time warping (DTW) distance and edit distance are two common shape-based distance metrics that can be used in these modified algorithms [78, 83]. Policker *et al*. applied fuzzy clustering methods to estimate the drift in the time series distribution and interpret the resulting matrix as weight in a time varying, mixture probability distribution function [151]. The model Policker *et al*. built has the capability of describing the changes in continuous time series more naturally and accurately [151]. The third approach is to use multi-step clustering with inputs extracted from the multiple dimensions of time series. In [152], time series are clustered in two steps–first grouped based on similarity in time, then further grouped based on shape similarity.

Discovering patterns from clinical time series is essential for early detection of illness, characterization of the severity of chronic conditions, and monitoring patient's recovery process. However, longitudinal clinical data are vastly underused due to the challenges involved in analyzing irregular time series. Past studies have attempted to capture trends from time series data using predefined features combined with a Gaussian mixture model for clustering episodes from the MIMIC II database [153]. It shows success in "search by example" task but heavily relies on the manufactured features. In another study, patterns and features are drawn from physiological time series data using an unsupervised approach which converts continuous values to discrete symbols and builds "bag-of-words" representations of time series [154]. A more recent work leverages a switching latent "topic" to construct informative features to rich and heterogeneous data generated in the ICUs [155]. These studies all have shown relative successes in pattern recognition from clinical time series data, but they rely on the availability of complete, high-

frequency bedside data, which are not often available in the case of laboratory measurements and vital signs.

In this chapter, several distance (similarity) metrics and clustering methods are leveraged in analysis of sparse, irregular clinical laboratory measurement time series data. A new approach is developed to cluster irregular (with unequal lengths and uneven intervals) Multivariate Time Series (MTS). In this chapter, the goal is to use these laboratory measurement data to characterize treatment outcomes.

5.2 MATERIAL AND METHODS

5.2.1 Distance Metrics and Clustering Methods for Univariate Time Series Data

For univariate clinical laboratory time series, this thesis experiments with DTW [78], Time Warp Edit Distance (TWED) [76], and a zero-padding periodogram discrepancy statistic (ZPDS) [156] similarity metrics. DTW is a shape-based distance metric which can be effective in detecting the different trends of time series. TWED benefits from the triangular inequality property and can match time series with time shifting tolerance. Consequently, both DTW and TWED are used to detect the change or trend in patient laboratory measurements overtime. In addition, both are elastic measures, meaning that they can be applied to time series with unequal length and uneven intervals.

Let $\{x_t, t = 1, \dots, n_x\}$ and $\{y_t, t = 1, \dots, n_y\}$ be two time series of different sizes. Without loss of generality, assume that $n_x > n_y$, we extend the shorter series by adding zeros and getting a new series y'_t. The ZPDS similarity metric between these two time series is defined by

$$\mathbf{d_{ZP}}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{\mathbf{1}}{\boldsymbol{m_x}} \sum_{\mathbf{j=1}}^{\mathbf{m_x}} [\mathbf{P_x}(\boldsymbol{\omega_j}) - \mathbf{P_{y'}}(\boldsymbol{\omega_j})]^2}, \qquad \textbf{(1)}$$

Where $P_x(\omega_j)$ and $P_{y'}(\omega_j)$ are the periodograms of series x_t, and y'_t, respectively. The periodogram of series x_t is given by

$$\mathbf{P_x(\omega_j) = \frac{1}{n_x}\left|\sum_{t=1}^{n_x} x_t e^{-it\omega_j}\right|^2}, \quad \textbf{(2)}$$

Where $\omega_j = 2\pi j/n_x$, for $j = 1, \ldots, m_x$, with $m_x = [n_x/2]$, the largest integer less or equal to $n_x/2$. [156].

To cluster time series data based on these metrics, hierarchical or medoid-based clustering algorithms should be applied because it is difficult to determine the length of the cluster centers when using partition-based clustering algorithms such as K-means [87]. In this chapter, agglomerative clustering [157] and K-medoids [158] clustering algorithm are applied on the pairwise distance matrix calculated using TWED and ZPDS metrics. Pre-processing techniques including resampling and interpolation methods are also leveraged to transform time series so that conventional K-means algorithm can be applied.

5.2.2 Distance Metrics and Clustering Methods for Multivariate Time Series Data

This thesis proposes a novel clustering approach for irregular MTS based on existing distance metrics for variable-length time series. DTW, soft-DTW, and GAK can be used to calculate the pairwise distances between variable-length univariate time series. An aggregation function is then applied to the distance between all pairs of corresponding univariate time series composing the MTS. This produces a pairwise distance matrix representing the similarity between each pair of patients. GAK [84] can be used to quantify the similarity between two time series of variable lengths. It is positive definite, fast to compute, and operates on the whole spectrum of costs of alignments and thus contains a richer statistic than DTW, which considers only the minimum of

the set of costs [84]. GAK distance is equal to the sum of the exponentiated and sign changed similarities of every alignment pairs:

$$\mathbf{GAK}(\mathbf{x},\mathbf{y}) = \sum_{\boldsymbol{\pi}\boldsymbol{\epsilon}\mathbf{A}(\mathbf{n},\mathbf{m})} \prod_{\mathbf{i}=\mathbf{1}}^{|\boldsymbol{\pi}|} \boldsymbol{\kappa}\left(\mathbf{x}_{\boldsymbol{\pi}_1(\mathbf{i})}, \mathbf{y}_{\boldsymbol{\pi}_2(\mathbf{i})}\right), \qquad \textbf{(3)}$$

where $A(n, m)$ is the set of all possible alignments between two series of length n and m, and any alignment pair (π_1, π_2) satisfies the warping restriction $\left(\frac{\pi_1(i+1)-\pi_1(i)}{\pi_2(i+1)-\pi_2(i)}\right) \epsilon \left(\frac{0}{1}\right), \left(\frac{1}{0}\right), \left(\frac{1}{1}\right)$[84]. Here, κ is a positive definite kernel function, and the Gaussian Kernel is used. Distance between each pair of MTS is then calculated by applying an aggregation function on the GAK distance between each pair of corresponding univariate time series. Here, we aggregate the distances using the average function. Specifically, given two patients P^1 and P^2 each characterized by m laboratory time series $P_1^1 P_2^1 \ldots P_m^1$, $P_1^2 P_2^2 \ldots P_m^2$, respectively, the aggregated distance is equal to

$$\frac{\mathbf{1}}{\mathbf{m}} \sum_{\mathbf{t}=\mathbf{1}}^{\mathbf{m}} \mathbf{GAK}(\mathbf{P_t^1 P_t^2}). \qquad \textbf{(4)}$$

Alternatively, weighted average, median, or the sum function could be used as the aggregation function depending on which laboratory time series is (are) more important. The aggregated distance represents an alignment score over each pair of univariate time series and provides a holistic similarity measure for the pair of MTS. In the case where patients are characterized by different types of laboratory time series, the intersection of these laboratory measurements should be selected to form MTSs consisting of the same number of laboratory time series.

Similar to univariate time series data, K-medoids clustering is applied on the pairwise distance matrix calculated using the aggregated GAK distances.

5.3 RESULTS

5.3.1 Clustering Patients with Diabetes Using Glycated Hemoglobin (HbA1c) Lab Measurements

The CCI-health database contains 267,666 diabetic patients. For each patient, treatment duration is determined by calculating the elapsed time between diagnosis (indicated by the first prescription of a diabetic medication) and the last recorded activity (i.e. procedure, laboratory, etc.). To characterize treatment outcome quality for these diabetic patients, Glycated hemoglobin (HbA1c) lab measurement recorded throughout the treatment duration are used as indicators of treatment outcome. In this analysis, only patients with 7 or more HbA1c measurements recorded are included in the dataset. This resulted in a total of 3,875 patients. A sliding window with a size of five is performed on each patient's HbA1c measurements, to reduce the noise in the time series. The 3,875 patients are clustered into two outcome groups based on their smoothed HbA1c lab measurement series. Distances between each pair of time series is calculated using the ZPDS metric. Finally, agglomerative clustering with average linkage is applied to the calculated pairwise distance matrix to cluster these patients into two groups. As a result, 3,475 patients are clustered into the "good outcome" group, and the remaining 400 patients are clustered in the "medium outcome" group. The outcome quality of the two groups are characterized as "good" versus "medium" based on the overall trend of HbA1c lab measurements in each group (Figure 5.1). It is evident that the medium outcome group has higher average HbA1c results and more fluctuations.

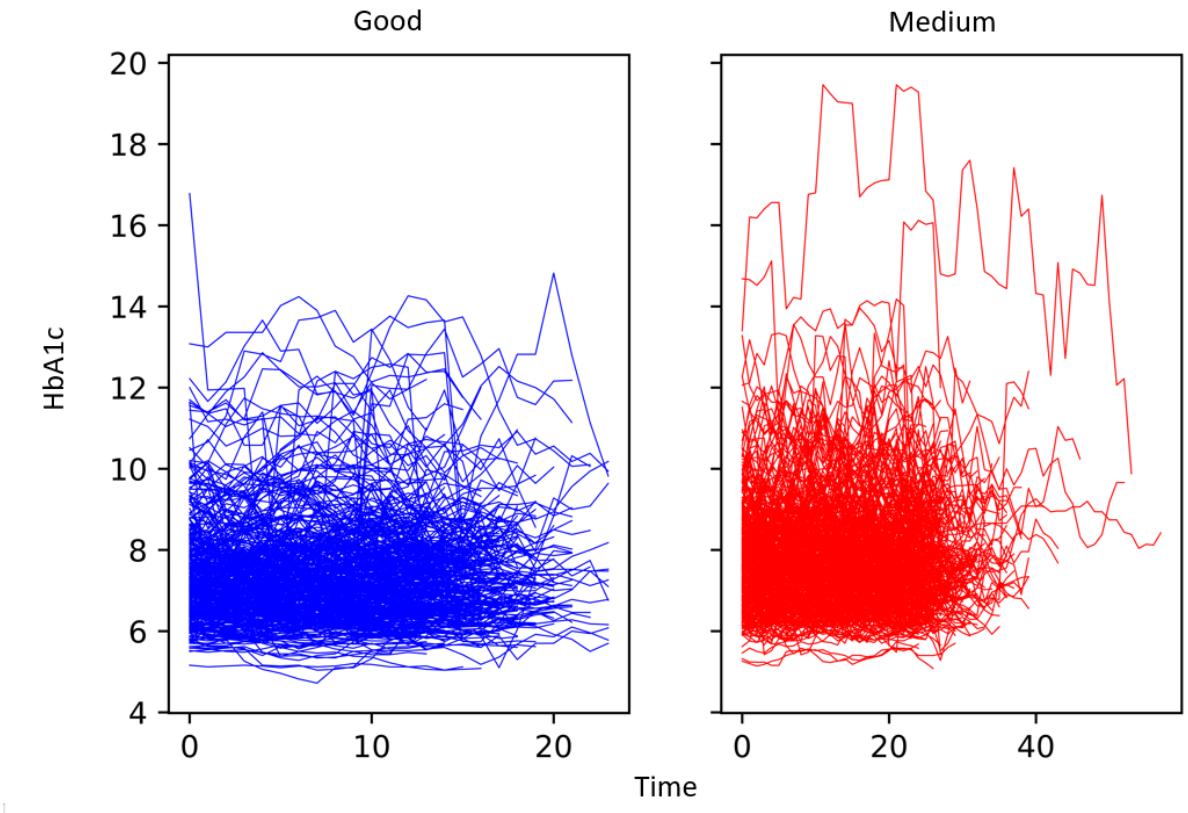


Figure 5.1. Agglomerative clustering with ZPDS distance for patients with diabetes based on HbA1c measurements.

5.3.2 Clustering Prostate Cancer Patients Using PSA Measurements

After all available PSA measurements are retrieved from the EPIC EHR, each patient's PSA time series is resampled to quarterly frequency (one measurement every three months). Gaps in the data are filled by propagating the non-NaN values forward first, and then backward along a series. This interpolation approach assumes that patients are expected to take new measurements when some progress during the treatment process are expected. Patients with very few (<5) measurements after resampling are removed from the dataset, resulting in a dataset containing 1,296 patients. The pairwise distance matrix is calculated using the TWED distance metrics, and K-medoids is applied to cluster these patients into three groups using the distance matrix. Figure 5.2 shows the PSA

values over the resampled periods for each of the three resulting patient clusters. The three clusters represent “Good”, “Worse”, and “Medium” outcome, respectively. It is notable that in the “Good” outcome group, the majority of patients’ PSA measurements are consistently dropping (showing that the machine’s concept of good outcome reflects the clinical criteria), whereas the majority of the PSA measurements of patients in the “Worse” outcome are either rising or fluctuating within a high range. The “Medium” outcome group are characterized by PSA values that are fluctuating within a medium to low range but relatively consistent when compared to the “Worse” outcome group.

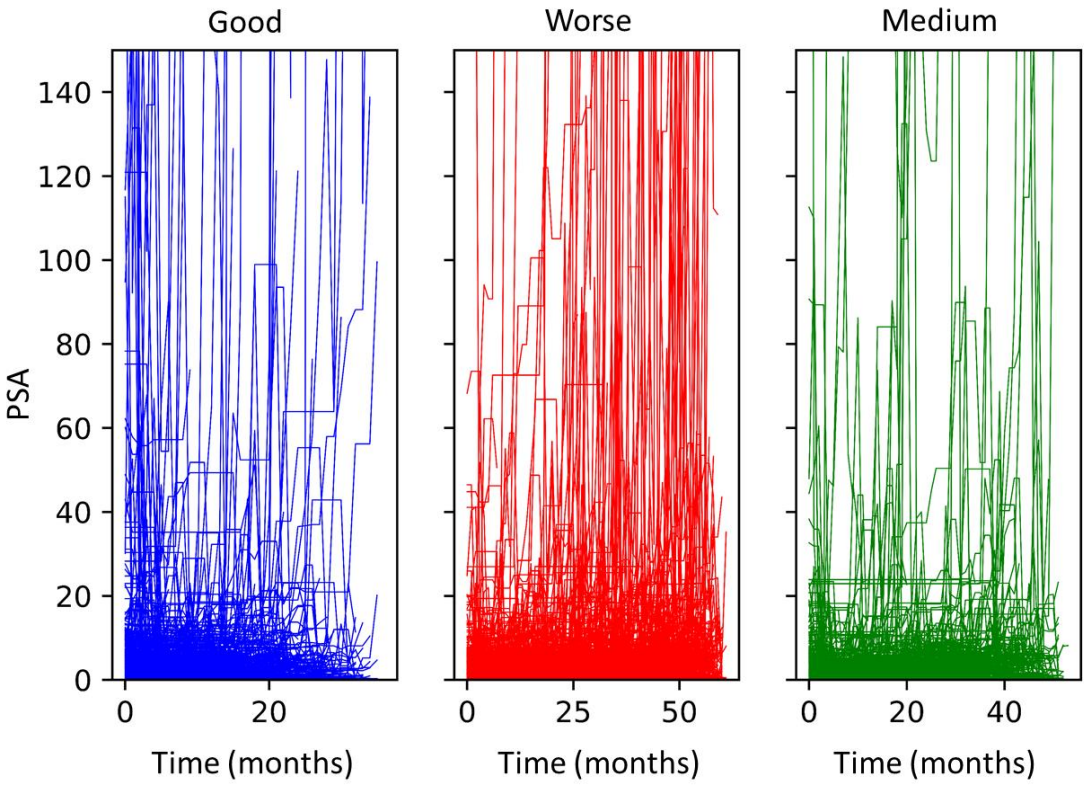


Figure 5.2. K-medoids clustering with TWED distance for patients with prostate cancer based on PSA measurements.

5.3.3 Clustering Patients with Chronic Kidney Disease Using eGFR Measurements

After extracting all eGFR measurements from the EPIC EHR, the distribution of the duration between first and last eGFR measurements for the entire patient cohort is examined. The longest duration lasted for 125 months while on average, patients' screening and treatment process lasted for about 120 months. eGFR measurements are resampled to a monthly frequency and linear interpolation is applied to fill in missing values. The first 120 data points are sampled from each patient's interpolated lab records so that each time series is equal in length. This sample period covers the treatment duration of most CKD patients. K-means algorithm with the Euclidean distance metrics is then applied to cluster these patients into 3 groups.

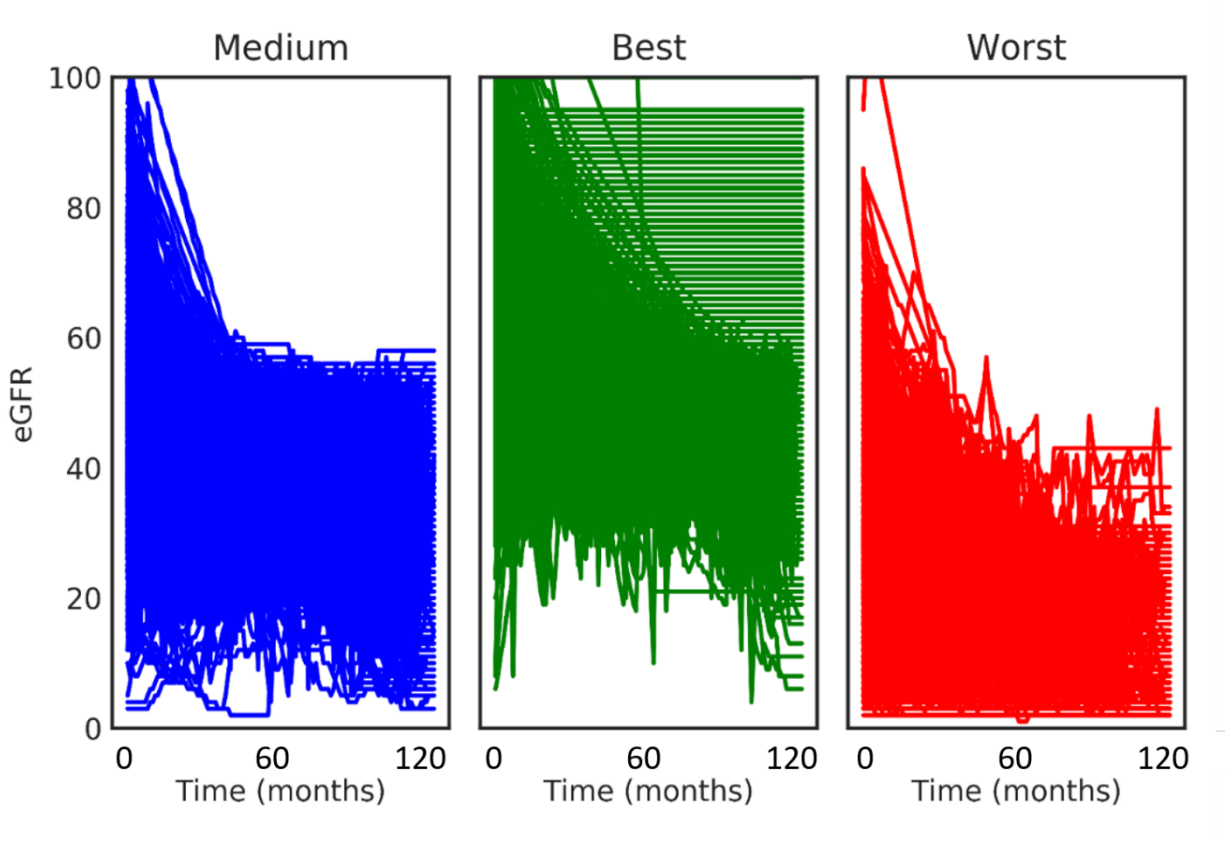


Figure 5.3. K-means clustering with Euclidean distance for patients with CKD based on eGFR measurements.

Figure *5.3* shows the eGFR values over the sampled 120 months for each of the three resulting patient clusters. The three clusters represent "Medium", "Best", and "Worst" treatment outcome, respectively. These outcomes are defined based on how fast eGFR declines, which represent the rate the patient loses his or her kidney function. It is evident that these three patient clusters differ in both the rate of eGFR decline and the ending eGFR levels (Table 5.1).

TABLE 5.1. RATE OF EGFR DECLINE AND ENDING LEVELS BY PATIENT CLUSTERS.

	eGFR decline rate	**eGFR ending level**
"Medium" outcome	Medium	Medium
"Best" outcome	Slow	High
"Worst" outcome	Fast	Low

5.3.4 Clustering Patients with Cardiovascular Disease Using Multiple Laboratory Measurements

The CCI-health database [6] contains 37,742 patients with CVD from 737 clinical sites. After pre-processing, each patient is characterized by 11 raw features including demographics, treatment duration, and co-existing conditions, and 1,757 standardized features in SNOMED-CT terminology including laboratory tests, diagnosed problems, and medications. For each patient, treatment duration is determined by calculating the elapsed time between diagnosis (indicated by the first prescription of a medication) and the last recorded activity (i.e. procedure, lab, etc.). Measurements of lipids and lipoproteins are processed into time series, since these are closely related to cardiovascular conditions and can potentially be used to characterize the severity of CVD. Lack of high-density lipoproteins (HDL) is significantly associated with the development of coronary heart disease [159]. In contrast, low-density lipoprotein increases the risk of heart disease and is considered a "bad" cholesterol [159]. Triglycerides are also associated with incidence of heart disease but has a less significant effect [159]. They are not directly atherogenic but represent an important marker of CVD risk [160].

In the analysis, HDL, LDL, and Triglycerides measurements are combined to form an MTS containing three time series for each patient used for clustering. Each of these time series are resampled to quarterly frequency. Gaps in the data are filled by propagating the non-NaN values forward first, and then backward. For each of the three types of laboratory measurements, patients with less than 3 measurements after resampling from the dataset are removed. This produces a dataset containing 450 remaining patients. The GAK distance between each pair of corresponding time series is calculated. The pairwise distance between each pair of MTS is then obtained by averaging the three distances for each pair of corresponding univariate time series. K-medoids

clustering is performed on the final distance matrix, partitioning the patients into three groups. The quality of clusters is evaluated both visually and quantitatively. Visually, trends of laboratory measurements are shown with boxplots of each patient's measurement taken at each time point. Quantitatively, the following statistics are calculated for each cluster: 1) median of first measured value; 2) median of the last measured value; 3) difference between the two medians. Since the goal is to segregate patients with different treatment outcomes, ideal clusters of patients should exhibit different trends of lab measurements. The three clusters of patients are characterized as having "Good", "Medium", and "Worse" outcomes by comparing the trends of laboratory measurements based on the visualizations and their three summary statistics.

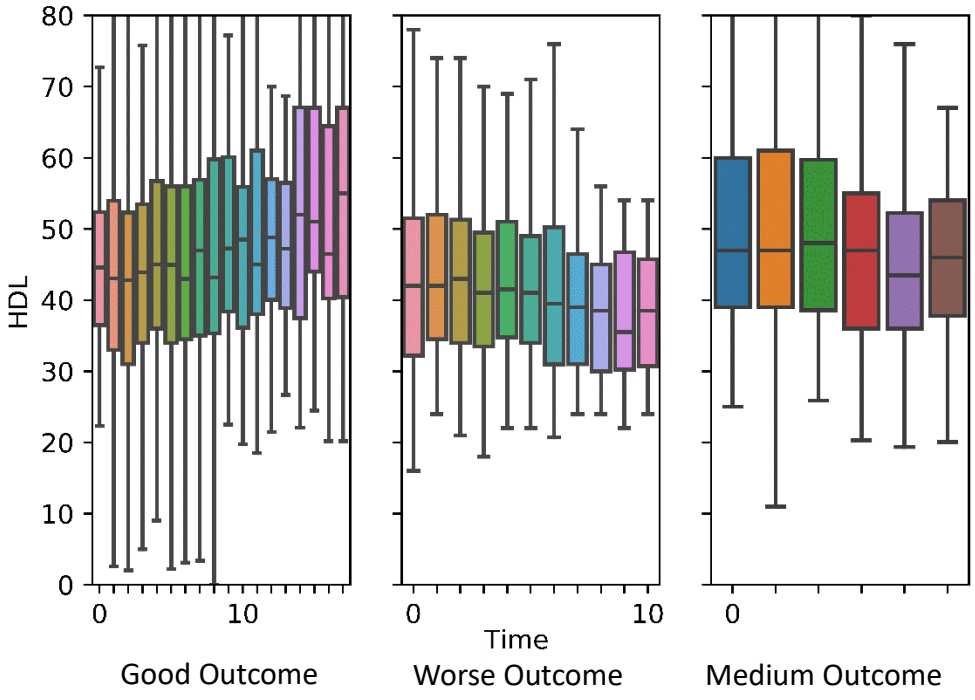


Figure 5.4. Boxplots of HDL laboratory records at each time point by cluster.

Figure 5.4, Figure 5.5, and Figure 5.6 show the boxplots of HDL, LDL, and Triglycerides laboratory records at each timepoint by cluster. Table 5.2 shows the summary statistics for each

laboratory measurement by cluster. The good outcome cluster shows an upward trend in HDL and downward trends in both LDL and Triglycerides (based on visualizations). It also has the largest decrease in median values from the first to the last measurement for both LDL and Triglycerides (based on summary statistics). The worse outcome group shows a downward trend in HDL and an upward trend in Triglycerides (based on visualizations) and has the largest increase in median values from the first to the last measurement for Triglycerides (based on summary statistics).

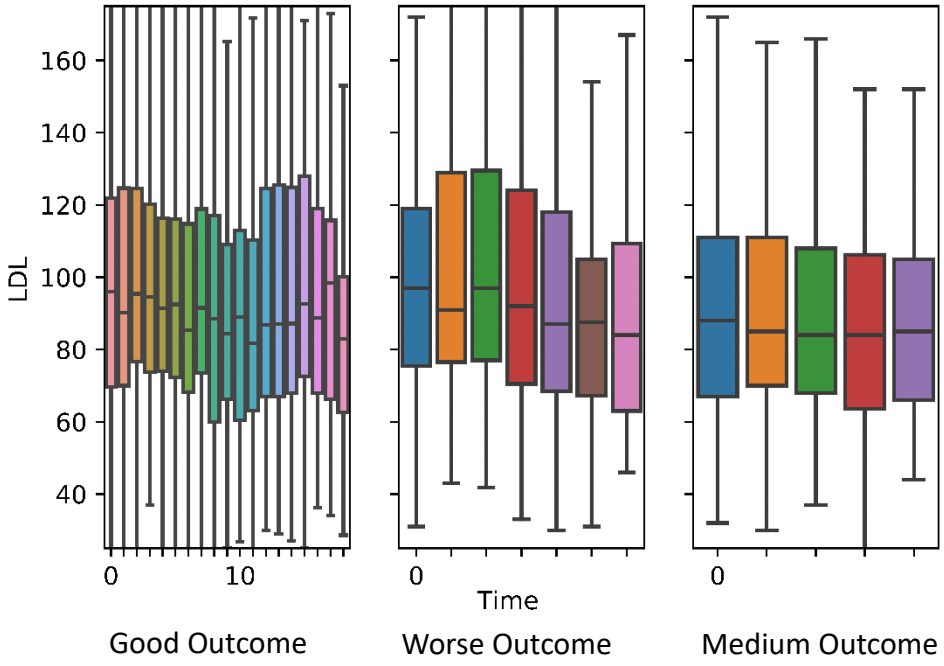


Figure 5.5. Boxplots of LDL laboratory records at each time point by cluster.

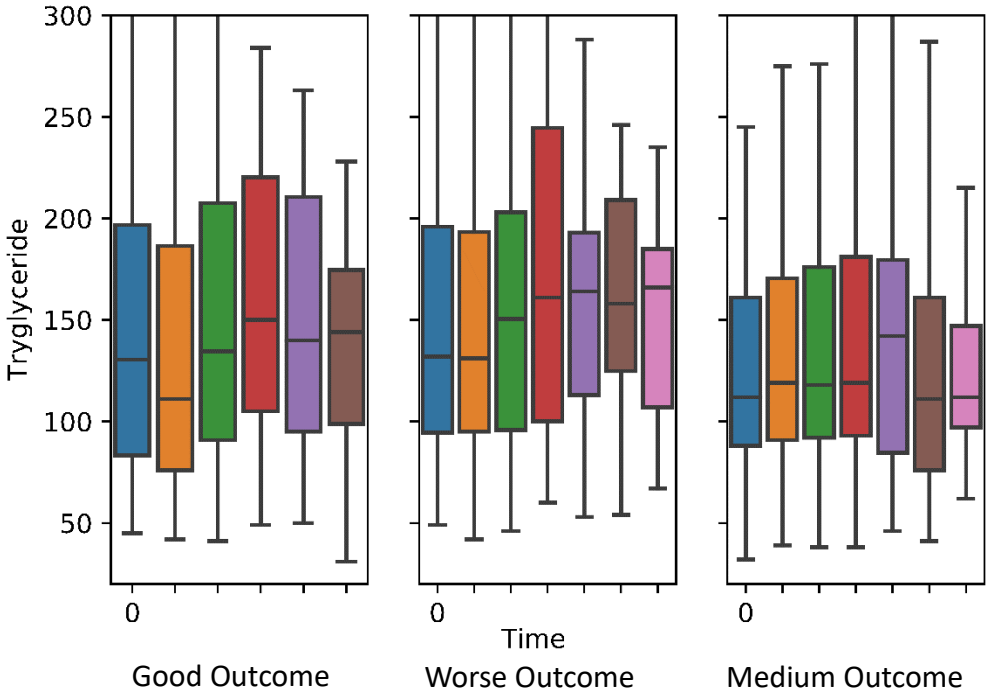


Figure 5.6. Boxplots of Triglycerides laboratory records at each time point by cluster.

TABLE 5.2. SUMMARY STATISTICS BY CLUSTER

	Median of	**HDL**	**LDL**	**Triglycerides**
Cluster 1 (Good Outcome)	First Value	44.6	96	130.5
	Last Value	45	86.5	125.5
	Difference	0.4	-9.5	-5
Cluster 2 (Worse Outcome)	First Value	42	97	132
	Last Value	43	93	156
	Difference	1	-4	24
Cluster 3 (Medium Outcome)	First Value	47	88	112
	Last Value	47	82	113
	Difference	0	-6	1

5.4 DISCUSSION

This chapter uncovers patient subgroups with varying quality of treatment outcomes based on longitudinal laboratory records measured during care delivery. Using time-series clustering, treatment outcomes of patients with diabetes, prostate cancer, CKD, and CVD are characterized based on the trend of HbA1c, PSA, eGFR, and cholesterol-related laboratory measurements,

respectively. Specifically, agglomerative clustering with ZPDS and K-medoids clustering with TWED distance are applied to univariate time series of varying lengths, and K-means clustering with Euclidean distance is applied to pre-processed univariate time series of equal lengths. In addition, a novel approach to cluster irregular MTS by aggregating the GAK distance measure over univariate time series is introduced. This allows us to characterize treatment outcome based on multiple laboratory records for each patient from a more holistic perspective. Each of these clustering approaches is able to create well-segregated clusters that differ in the trends of laboratory records.

By comparing these clustering approaches, it is evident that their strengths differ in categorizing different types of trends within the time series data. In particular, the ZPDS metric can differentiate between stationary and non-stationary processes [156]. TWED, on the other hand, can distinguish time series data with different shapes despite their absolute range of values. Euclidean distance used in the K-means clustering algorithm excels in segregating time series whose values differ greatly in magnitudes. Finally, GAK distance is used to characterize trends over multiple time series in the clustering process. Depending on types of disease and laboratory measurements, and the goal of clustering, different metrics should be leveraged to achieve the best clustering outcome.

CHAPTER VI

DISCRIMINATORY ANALYSIS FOR CLINICAL PROCESS AND OUTCOME IMPROVEMENT

6.1 INTRODUCTION

Some of the causes of deficiencies in healthcare delivery in the United States include misuse of limited medical resources and staff time, practice variability across organizations, and the lack of best practice and outcome-driven standards. Data-driven healthcare has the potential to revolutionize care delivery and reduce costs. A major challenge is that providers must sift through and analyze intelligently mountains of disparate data to materialize the substantial gain. Machine learning provides a solution to this problem by efficiently establishing relationships between multiple features [161]. In the EHR, every patient instance is represented by a set of features. These features may be categorical (e.g. which type of drug is prescribed), continuous (e.g. age), or binary (e.g. whether a condition is diagnosed). Given certain labels for these patient instances, supervised learning models can be trained to predict the correct labels based on input features.

Feature, or variable selection methods is the process of fitting a supervised learning model using only a subset of features as predictors for an outcome [162]. It reduces model complexity and shortens training time, addresses the problem of the "Curse of Dimensionality" [163], avoids overfitting [164], and simplifies the model so that researchers and practitioners can interpret the results better [162]. Because clinical data are extremely complex and heterogeneous, it is essential to determine the appropriate forms of input variables to represent clinical features. For instance, a particular prescription can be represented as a binary variable indicating if a patient received the prescription, a nominal variable indicating the brand of the drug, or a continuous variable indicating the dosage.

By leveraging supervised learning models combined with feature selection, healthcare researchers and practitioners can 1) identify small sets of key factors that impact treatment outcome and optimize treatment regimen, 2) design personalized treatment tailored to individual patient needs, 3) enhance patients' experiences, and 4) reduce healthcare resources and costs.

In chapter V of this thesis, unsupervised learning models are built to identify patient subgroups characterized by differing trends of laboratory measurements. Patients in each cluster are labeled with a distinct treatment outcome based on numerical and visual analytics. Because these labels are discrete, classification models should be leveraged. In this chapter, various feature selection and classification models are used to build outcome prediction models. Patient demographics, treatment time and frequency, laboratory tests, vitals, co-existing conditions, medications, and clinical procedures are included as potential predictors of treatment outcome.

6.2 MATERIAL AND METHODS

In this chapter, discriminatory analysis is conducted on patient datasets labeled with varying treatment outcomes using various state-of-the-art supervised learning algorithms. The problem of class imbalance is addressed by adjusting the cost of misclassification by specifying different weights for each class. A customized scoring metric is also developed to select models with the highest accuracies across all classes. In addition, model performances are evaluated using the confusion matrix. Various feature selection methods are utilized to reduce the dimension of datasets and improve model performance and efficiency. The goal is to identify critical risks and treatment factors that impact treatment outcome.

6.2.1 Supervised Learning Models

Random Forest, Naïve Bayes, Support Vector Machine, Gradient Boosting, Logistic Regression are the main supervised learning models applied to various datasets. XGBoost [165] is also applied to classification tasks due to its high scalability for high-volume data. Different cost functions and training evaluation metrics are compared for different datasets. Multi-stage classification is performed to refine model performance.

6.2.2 Feature Selection Methods

Several popular feature selection methods including Randomized Lasso, feature importance score-based selection, and wrapper-based approaches such as recursive feature elimination [166] are applied to various datasets. The Hilbert-Schmidt Independence Criterion Lasso (HSIC Lasso) is also explored due to its ability to capture the non-linear relationship between input and output [167]. In addition, results from our in-house DAMIP/PSO machine learning framework [107] are compared against output from these commercialized algorithms.

6.2.3 Class-weighted Balanced Accuracy

In this thesis, a scoring metric for model performance is developed based on the balanced accuracy function [168] to address the issue of class imbalance. In the case of binary classification, the balanced accuracy is defined as $\frac{1}{2}(\frac{\mathrm{TP}}{\mathrm{P}}+\frac{\mathrm{TN}}{\mathrm{N}})$, where TP, P, TN, and N are true positive (the proportion of actual positives that are correctly identified as such), actual positives, true negative (the proportion of actual negatives that are correctly identified as such), and actual negatives [168]. This balanced accuracy scoring function takes into account the accuracy of both classes and ensures that a "good" classifier do not take advantage of imbalance classes [168]. In this thesis, a metric is developed based on balanced accuracy to further penalize classifiers with uneven performances for each class. It is defined as $\frac{\mathrm{TP}}{\mathrm{P}}\frac{\mathrm{P+N}}{\mathrm{P}}+\frac{\mathrm{TN}}{\mathrm{N}}\frac{\mathrm{P+N}}{\mathrm{N}}$. In this definition, each per-class

accuracy term from the balanced accuracy function is multiplied by the inverse of that class's proportion. It is given the name "class-weighted balanced accuracy". This function favors classifiers that achieves higher accuracies for the smaller class.

6.3 RESULTS

6.3.1. Patients with Prostate Cancer

Patient treatment outcome for the prostate cancer patient cohort is determined by the "disease-free" status based on PSA laboratory measurements. A Python script is developed to remove patients with less than 5 measured PSA records or without "Prostate Cancer" in their diagnosed problem lists. A patient is characterized as "disease-free" by the Python script if the three most recent PSA records all fall within the normal PSA range defined by one's age group (Table 6.1). the distribution of patients by outcome class is shown in Table 6.2.

TABLE 6.1. NORMAL PSA RANGE BY AGE GROUP.

Age Range	Normal PSA Range
age 49 or younger	0.0–2.5 ng/mL
age 50 to 59	0.0–3.5 ng/mL
age 60 to 69	0.0–4.5 ng/mL
age 70 or older	0.0–6.5 ng/mL

TABLE 6.2. SUMMARY OF PATIENT OUTCOME GROUP DISTRIBUTIONS.

Outcome	Not Disease-Free	Disease-Free
# of instances	263	1045

Classification models are built using the "disease-free" status as the outcome label. A total of 10,003 features are used as input features, including age, treatment procedures, diagnoses, labs, and medications during the treatment period (defined as the time between diagnosis of prostate cancer and the last measured PSA). All features are binary, with "1" representing "had received", and "0" representing "had not received", except for patient age, which is continuous. Three different feature selection methods are compared: HSIC Lasso, Randomized Lasso, and Recursive Feature Elimination with ExtraTrees Classifier [169]. Six different classification algorithms

including Naive Bayes, Random Forest, Logistic Regression, SVM, Gradient Boosting, and ExtraTrees are used. These algorithms are implemented with the Python scikit-learn package [170]. Each classification model is trained on 80% of (1,046) patients with 10-fold cross-validation and evaluated on 20% hold-out set of (262) patients. The two sets are partitioned using stratified random sampling. The confusion matrix is used as evaluation method. Overall, SVM achieves the best results when the "class_weight" parameter is set to "balanced" in the sklearn.svm.SVC class. The "scoring" parameter is set to "f1_weighted" for evaluation of model performances. These two parameter settings adjust for the imbalanced class issue by increasing the impact of the minority class. The RandomizedLasso module is used to select 29 discriminatory features including "History of PrCA", "Cancer metastatic to the bone", "TRANSPERINEAL PLMT NDL/CATHS PROSTATE RADJ INSJ" (Brachytherapy), "Zytiga", "Radiation Therapy", "Lipid Panel", "PSA", "Enzalutamide", "Creatinine", "Injection Docetaxel", etc. Confusion matrices are shown in Table 6.3 and Table 6.4, corresponding to cross-validation and blind prediction results.

TABLE 6.3. 10-FOLD CROSS-VALIDATION CONFUSION MATRIX (SCORING FUNCTION: "F1_WEIGHTED").

	Not disease-free	**Disease-free**
Not disease-free	158 (75.2%)	52
Disease-free	246	590 (70.6%)

TABLE 6.4. BLIND PREDICTION CONFUSION MATRIX (SCORING FUNCTION: "F1_WEIGHTED").

	Not disease-free	**Disease-free**
Not disease-free	42 (79.2%)	11
Disease-free	66	143 (68.4%)

While keeping all training steps, feature selection, and other parameters intact, the weighted-class balanced accuracy is also implemented and used as a "scoring" parameter to select the best performing model. The SVM model is chosen, which improves the blind prediction accuracy of

the disease-free class while keeping the similar accuracy for the not disease-free class (Table 6.5 and Table 6.6).

TABLE 6.5. 10-FOLD CROSS-VALIDATION CONFUSION MATRIX (SCORING FUNCTION: "CLASS-WEIGHTED BALANCED ACCURACY").

	Not disease-free	**Disease-free**
Not disease-free	150 (71.4%)	60
Disease-free	235	601 (71.9%)

TABLE 6.6. BLIND PREDICTION CONFUSION MATRIX (SCORING FUNCTION: "CLASS-WEIGHTED BALANCED ACCURACY").

	Not disease-free	**Disease-free**
Not disease-free	42 (79.2%)	11
Disease-free	59	150 (71.7%)

To refine these initial results, a second round of machine learning analysis is performed. First, the size of patient dataset is reduced by removing patients whose record do not contain the diagnosed problem "Malignant neoplasm of prostate". This new dataset contains 521 disease-free patients and 114 not disease-free patients. Additionally, six key treatment features are identified from features selected in the initial analysis: Laparoscopic Prostatectomy, CT Scan of Chest, Enzalutamide, Brachytherapy, Radiation Therapy, and Zoledronic Acid IVPB. The PSA laboratory value corresponding to the closest date before each of these six treatments is used to determine the stage of prostate cancer at which the patient was given the treatment. "0" represent "had not received treatment", and "1", "2", "3" corresponds to PSA values 0-10, 10-20, and 20 or more, respectively. These six new categorical input features are included as input features for each classification model. Each classifier is re-trained on 70% of (444) patients with 10-fold cross-validation and evaluated on 30% hold-out set of (191) patients. The two sets were re-partitioned using stratified random sampling. ExtraTrees classifier achieves the best performance, and the results are shown in Table 6.7 and Table 6.8. Four of the six newly created treatment features are

selected with HSIC Lasso along with 8 other discriminatory features. Further analysis and expert knowledge on these treatment features can provide us with insights on the optimal timing of treatment. For comparison, the feature selection and classification steps are repeated on this new patient dataset without including the six new features. With the same number of features (12) selected from the original feature set, the best performing ExtraTrees classifier achieves worse cross-validation accuracies for both groups. As for blind prediction, it loses nearly 15% accuracy for the not disease-free group, while only improving the disease-free group's prediction accuracy by 4.4% (Table 6.9 and Table 6.10).

TABLE 6.7. 10-FOLD CROSS-VALIDATION CONFUSION MATRIX WITH NEW TREATMENT FEATURES.

	Not disease-free	**Disease-free**
Not disease-free	57 (71.3%)	23
Disease-free	61	303 (83.2%)

TABLE 6.8. BLIND PREDICTION CONFUSION MATRIX WITH NEW TREATMENT FEATURES.

	Not disease-free	**Disease-free**
Not disease-free	29 (85.3%)	5
Disease-free	36	121 (77.1%)

TABLE 6.9. 10-FOLD CROSS-VALIDATION CONFUSION MATRIX WITHOUT NEW TREATMENT FEATURES.

	Not disease-free	**Disease-free**
Not disease-free	55 (68.7%)	25
Disease-free	79	285 (78.3%)

TABLE 6.10. BLIND PREDICTION CONFUSION MATRIX WITHOUT NEW TREATMENT FEATURES.

	Not disease-free	**Disease-free**
Not disease-free	24 (70.6%)	10
Disease-free	29	128 (81.5%)

6.3.2. Patients with Diabetes

Treatment outcome for the 3,875-patient cohort with diabetes is determined by clustering in chapter 5. Each patient is characterized by 24 raw features including hospital site, demographics, laboratory tests and results, prescriptions, treatment duration, chronic conditions, blood pressure, number of visits and visit frequencies, and 2,205 SNOMED-CT general standardized and

unambiguous concepts related to diagnosis, medication, and laboratory. These are included as input features in the classification models. The goal is to predict these patients' treatment outcome. The dataset is partitioned into a training set and an independent set for blind prediction using stratified random sampling. The training set consists of 2,325 patients (60% of the population), and the blind prediction set consists of 1,550 patients (40% of the population). Table 6.11 summarizes the number of patients in each set.

TABLE 6.11. PARTITIONING OF DIABETES TREATMENT OUTCOME CLUSTERS FOR CLASSIFICATION ANALYSIS.

	10-fold Cross Validation Training Set			**Blind Prediction Set**		
Total Patients	Total	Good outcome	Medium outcome	Total	Good outcome	Medium outcome
3875	2325	2074	251	1550	1401	149

In the classification analysis, models are first applied to the training set to establish the classification rules using 10-fold cross-validation unbiased estimate. The predictive accuracy of each classifier is further assessed by applying the rule to the blind prediction set.

To demonstrate the effectiveness of standardized clinical terminologies, classification models built from two input feature sets are compared: the first set of input features included only the 24 raw features; the second set included the additional 2,205 generalized SNOMED-CT concepts obtained from terminology mapping.

The DAMIP feature selection and classification framework [107] is contrasted with logistic regression, naïve Bayes, radial basis function networks (RBFNetwork), Bayes Net, J48 decision tree, and sequential minimal optimization (SMO) approaches implemented in Weka 3.6.13. In these Weka classifiers, feature selection is performed using the "InfoGainAttributeEval" as the Attribute Evaluator and "Ranker" as the Search Method to select at most 200 features.

Table 6.12 contrasts our in-house DAMIP classification results with six Weka classifiers using 24 raw features. Uniformly the six classifiers suffer from group imbalance and tend to place all patients into the larger "good outcome" group. In contrast, the DAMIP classifier selects 5 discriminatory features among the 24 and is able to achieve high classification and blind prediction accuracies for both groups. (We remark that the commonly used Pap Smear diagnosis test has an accuracy of 70%).

TABLE 6.12. COMPARISON OF DAMIP RESULTS AGAINST OTHER CLASSIFICATION METHODS.

Classifier	**Features set**	**10-fold Cross Validation Accuracy**			**Blind Prediction Accuracy**		
		Overall	Good outcome	Medium outcome	Overall	Good outcome	Medium outcome
Logistic regression	treatment duration, visit frequency, hypertension, hyperlipidemia, cvd, stroke, emphysema, asthma, race, gender, age, height, weight, patient site, 5 systolic blood pressure measurements, and 5 diastolic blood pressure measurements	89.68%	96.87%	30.28%	90.77%	97.14%	30.87%
Naïve Bayes		87.05%	92.77%	39.84%	87.94%	92.72%	42.95%
RBFNetwork		88.22%	98.36%	4.38%	89.03%	97.22%	12.08%
Bayes Net		87.66%	92.96%	43.82%	87.23%	91.86%	43.62%
J48		89.16%	97.69%	18.73%	90.65%	97.72%	24.16%
SMO		89.16%	99.42%	4.38%	90.77%	99.64	7.38%
DAMIP (5 features)	treatment duration, visit frequency, hyperlipidemia, asthma, provider site	**80.1%**	**80.6%**	**75.1%**	**81.9%**	**81.9%**	**81.3%**

TABLE 6.13. COMPARISON OF PREDICTION ACCURACIES WITH ADDITIONAL 2,205 MEDICAL TERMINOLOGIES OBTAINED FROM MAPPING.

Classifier	**Features set**	**10-fold Cross Validation Accuracy**			**Blind Prediction Accuracy**		
		Overall	Good outcome	Medium outcome	Overall	Good outcome	Medium outcome
Logistic regression	treatment duration, patient site (practice), height, asthma, visit frequency, race, hypertension, 2 diastolic blood pressure measurements, and 191 new features from terminology mapping	87.83%	94.07%	36.25%	89.81%	94.93%	41.61%
Naïve Bayes		74.75%	76.81%	57.77%	75.29%	77.73%	52.35%
RBFNetwork		88.73%	99.42%	0.40%	90.39%	100.0%	0.00%
Bayes Net		80.34%	83.37%	55.38%	80.84%	83.73%	53.69%

TABLE 6.13. continued

J48		89.12%	96.62%	27.09%	87.94%	95.07%	20.81%
SMO		89.25%	96.77%	27.09%	90.52%	97.29%	26.85%
DAMIP	treatment duration, visit frequency, asthma, + 6 features from mapping: Injection of therapeutic agent (procedure), Oral form levofloxacin (product), Clopidogrel (product). Aspirin (product), Nystatin (product), and Metformin (product)	**88.7%**	**88.6%**	**89.5%**	**87.9%**	**86.7%**	**89.7%**

Table 6.13 shows classification results after including 2,205 additional features for each patient. With these added features, DAMIP improves its prediction accuracies of both the "good outcome" group and the "medium outcome" group. It is observed that instead of selecting the "provider site" as one of the discriminatory features, DAMIP selects the type of procedures and medications used instead. Identifying these features facilitates dissemination of best practice and target treatment regime to specific patients among the sites. For the six Weka classifiers, the results improve only slightly on the "medium outcome" group.

6.3.3. Patients with Cardiovascular Disease

Treatment outcomes of CVD are determined for 450 patients using MTS clustering on their HDL, LDL, and Triglycerides laboratory measurements. Table 6.14 shows the distribution of these patients by CVD treatment outcome clusters.

TABLE 6.14. DISTRIBUTION OF PATIENTS WITH CVD BY OUTCOME CLUSTERS.

# of Patients	**138**	**233**	**79**
Outcome	Good	Medium	Worse

Multi-class classification models are built using these three clusters. Among the 1,768 input features (1757 generalized SNOMED-CT concepts and 11 raw features), most are binary with "1"

representing “had received”, and “0” representing “had not received”. For gender, “1” represents male, and “0” represents female. Age and treatment length are the only continuous input features. Two feature selection methods are compared: Randomized Lasso and Recursive Feature Elimination with Random Forest Classifier [169]. Eight classification algorithms implemented with the Python Scikit-learn package [170] including Logistic Regression, SVM, K-nearest neighbors, ExtraTrees, Random Forest, Decision Tree, Neural Network, Gradient Boosting, and Bernoulli Naïve Bayes are applied to the dataset. Each classifier is trained on 80% of (360) patients with 10-fold cross-validation and evaluated on a 20% hold-out set of (90) patients. The two sets are partitioned using stratified random sampling. Overall, ExtraTrees classifier achieves the best results. Recursive feature elimination with random forest selects 25 discriminatory features including treatment length, “surgical pathology procedure (procedure)”, “urine screening (procedure)”, “electrocardiogram finding (finding)”, “immunologic procedure (procedure)”, etc. Confusion matrices are shown in Table 6.15 and Table 6.16, corresponding to cross-validation and blind prediction results.

TABLE 6.15. 10-FOLD CROSS-VALIDATION CONFUSION MATRIX.

	Good	**Medium**	**Worse**
Good	92 (82.9%)	8	11
Medium	10	165 (88.7%)	11
Worse	6	16	41 (65.1%)

TABLE 6.16. BLIND PREDICTION CONFUSION MATRIX.

	Good	**Medium**	**Worse**
Good	24 (88.9%)	0	3
Medium	2	43 (91.5%)	2
Worse	1	6	9 (56.3%)

Of the three outcome groups, the classifier achieves high accuracy on both the “medium” and “good” outcome groups. The “worse” outcome group has many misclassified examples due to its small sample size. However, it is noticeable that many patients from the “worse” outcome group are misclassified into the “medium” outcome group. After re-examining the lab records for patients

in each cluster, the “medium” and “worse” outcome clusters are combined into one “Not Satisfactory” outcome cluster. A second round of classification is then performed using this new two-cluster dataset.

Each classifier is re-trained on 70% of (315) patients with 10-fold cross-validation and evaluated by a 30% hold-out set of (135) patients. The two sets are re-partitioned using stratified random sampling. Overall, Random Forest achieves the best results. Randomized Lasso selects 16 discriminatory features including ‘Benign neoplastic disease (disorder)’, ‘Disorder of lipoprotein AND/OR lipid metabolism (disorder)’, ‘Urine screening (procedure)’, age, treatment length, ‘Chronic disease of immune system (navigational concept)’, ‘Surgical pathology procedure (procedure)’, ‘Open wound of trunk (disorder)’, ‘Immunologic procedure (procedure)’, ‘Hormone measurement (procedure)’, ‘Chemical categorized structurally (substance)’, ‘Neoplasm by body site (disorder)’, ‘Screening finding (finding)’, ‘Inflammatory disorder of head (disorder)’, ‘Arteriosclerotic vascular disease (disorder)’, and ‘Hematopoietic system finding (finding)’. Confusion matrices are shown in Table 6.17 and Table 6.18.

TABLE 6.17. 10-FOLD CROSS-VALIDATION CONFUSION MATRIX.

	Good	**Not Satisfactory**
Good	76 (78.4%)	21
Not Satisfactory	14	204 (93.6%)

TABLE 6.18. BLIND PREDICTION CONFUSION MATRIX.

	Good	**Not Satisfactory**
Good	35 (85.4%)	6
Not Satisfactory	9	85 (90.4%)

Among the 16 selected features, 'Urine screening (procedure)’, ‘Benign neoplastic disease (disorder)’, ‘Chronic disease of immune system (navigational concept)’, ‘Surgical pathology procedure (procedure)’ are the most significant factors based on two-sample t-tests. Figure 6.1 shows the distribution of these four binary features by outcome group.

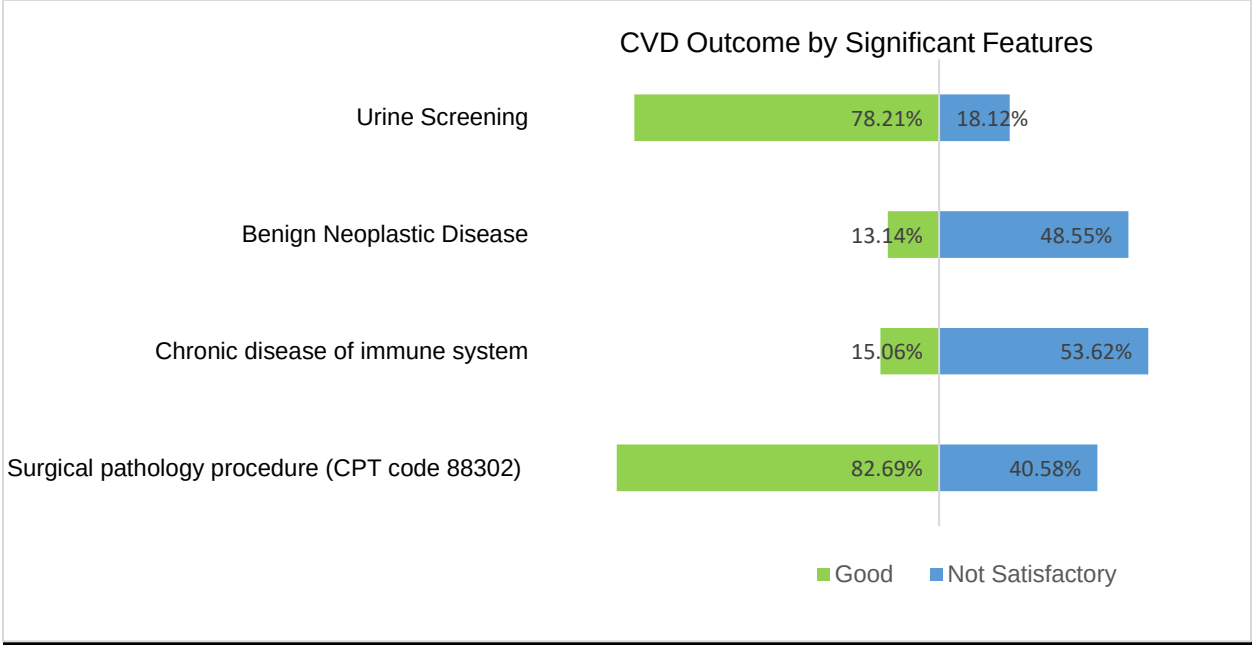


Figure 6.1. CVD outcome by four selected features

Using only these four features alone, above 70% prediction accuracy can be achieved for both patient groups (Table 6.19 and Table 6.20).

TABLE 6.19. 10-FOLD CROSS-VALIDATION RESULTS USING FOUR PRE-SELECTED FEATURES.

	Worse	**Satisfactory**
Worse	77 (79.8%)	20
Satisfactory	64	154 (70.7%)

TABLE 6.20. BLIND PREDICTION RESULTS USING FOUR PRE-SELECTED FEATURES.

	Worse	**Satisfactory**
Worse	33 (80.7%)	8
Satisfactory	17	77 (81.4%)

6.4 DISCUSSION

In this chapter, feature selection and classification models are developed to uncover patient and practice characteristics that can predict treatment outcome. Three cohorts of patients with prostate cancer, diabetes, and CVD respectively, are studied.

Treatment outcome for prostate cancer is determined with patients' most recent PSA laboratory measurements. The SVM classifier achieves 79.2% and 68.4% blind prediction

accuracies for the two outcome groups with 29 discriminatory features. Constructing more refined features indicating the timing (stage) of key treatment features significantly improves model performances. Table 6.21 shows the how the four features related to timing of treatment selected by HSIC Lasso impacts treatment outcome. It is notable that all four types of treatments result in better outcome if given at earlier stages of prostate cancer. Furthermore, Brachytherapy stands out as an effective means of treatment for majority of patients. These features provide key insights to optimal timing of prostate cancer treatment. However, until the impact of early intervention for prostate cancer is fully understood, particularly for patients of older ages, it is necessary to take care in choosing between early definitive therapy and active surveillance [171].

TABLE 6.21. PROSTATE CANCER PATIENT OUTCOME BASED ON TIMING OF KEY TREATMENT FEATURES.

Disease Free	Disease Free	No Treatment	Stage 1	Stage 2	Stage 3	Total
Xtandi (Enzalutamide)	No	83	6	7	18	114
	Yes	509	7	3	2	521
Brachytherapy	No	106	2	4	2	114
	Yes	289	188	40	4	521
CT Scan	No	65	22	5	22	114
	Yes	457	57	4	3	521
Zoledronic	No	89	4	4	17	114
	Yes	511	6	0	4	521

Treatment outcome for diabetes is determined by clustering using patients' HbA1c laboratory test results. Compared to other classification models, the DAMIP model is able to select a small set of discriminatory features to establish good classification rules. Specially, the DAMIP-rules can distinguish patients with good outcomes from those with medium outcomes with over 80% blind predictive accuracy. By including medical concepts obtained from terminology mapping, predictive capabilities of classification models are further improved, creating a more powerful evidence-based best practice discovery system. The features identified, including

treatment duration, frequency, co-existing condition, and type of regimens allow for design of "best practice" clinical practice guidelines.

Treatment outcome for CVD is determined via clustering using patients' HDL, LDL, and Triglycerides laboratory test results. The ExtraTrees classifier achieves nearly 90% blind prediction accuracy for both "Good" and "Medium" outcome groups with 25 features. Although the blind prediction accuracy for the "Worse" outcome group is only 56%, most of these patients are mis-classified as having "Medium" outcome. By redefining the "Medium" and "Worse" outcome groups into one "Not Satisfactory" group, the Random Forest classifier can achieve above 85% blind prediction accuracy for both the "Good" and "Not Satisfactory" outcome groups using only 16 features.

The next step is to use the discriminatory features identified and the criteria developed via the machine learning framework to design and optimize evidence-based treatment plans for each type of disease and to disseminate such knowledge through "rapid learning" across multiple sites. The ultimate goal is to increase quality and timeliness of care and maximize outcome and service performance of the healthcare system.

One limitation of this work is the fact that only a small portion of patients are selected for each cohort for discriminatory studies due to missing or low-quality data. Although the resulting classification models achieve high prediction accuracies, the discriminatory features can only serve as critical treatment or risk factors for this small subset of patients. Further studies should be done to better understand the significance of these features for the larger patient population through incorporating more patient samples in the analysis.

CHAPTER VII

REDUCING HEALTHCARE DISPARITY WITH REMOTE PATIENT MONITORING AND TELEHEALTH

7.1 INTRODUCTION

With an aging population and the continuous rise of healthcare costs, more affordable and accessible care must be explored. Healthcare expenditures in the United States totaled $3.8 trillion in 2014 [172], with roughly 90% spent on chronic illness [173]. It was estimated that about half of all adults—117 million —suffer from one or more chronic health conditions [174]. Specifically, in a 2011 study [175], it was estimated that adults with arthritis, diabetes mellitus, heart disease, and hypertension had average out-of-pocket spending of $1,814 per year. Among those with combinations of 2 or 3 conditions, average out-of-pocket spending ranged from $1,134 to $1,760. This contrasts to $795 per person for the overall population, and $343 for individuals without any chronic illness [174]. Reducing the occurrence of chronic conditions and the number of necessary healthcare provider visits, and timely access to medical consultations can help to reduce costs and improve health.

7.1.1 Current State of Telemedicine

The success of telemedicine can be proven through positive testimonies. A pilot trial at Partners HealthCare using remote monitoring for patients with congestive heart failure reduced hospital readmission by 9% compared to usual care [176]. Another home telehealth program introduced by the Veterans Health Administration (VHA) reported high satisfaction levels among 17,025 participating patients with one or more chronic illnesses [177]. It reduced number of bed days of care by 25% and the number of hospital admissions by 19 % compared to usual care [177]. Many telemedicine programs have also achieved significant cost savings. Established in 2001, The

Florida Initiative in Telehealth and Education (FITE) provided telemedicine clinics supplemented by online education for children with diabetes. The program saved $27,860 per year and reduced average hospitalizations from 13 to 3.5 per year [178]. A controlled study of eleven nursing homes suggests that by fully engaging in telemedicine, a nursing home could achieve a net savings of roughly $120,000 per year [179]. Moreover, telemedicine programs have exerted positive environmental impact. A study on the telemedicine consultation database at the University of California Davis Health System shows a 278-mile per consultation trip, which resulted in a total emissions savings of 1969 metric tons of CO2, 50 metric tons of CO, 3.7 metric tons of NOx, and 5.5 metric tons of volatile organic compounds between July 1996 and December 2013 [180].

While benefits and challenges for telehealth can be generalized to a certain degree nationwide, state statistics and trends can also provide much further detail and a more accurate understanding. 108 of Georgia's 159 counties were found to have persistent poor children's health and working age adult health. Compared to the urban area, Rural Georgia has a poorer, sicker, and older population, making it a critically important area to the state's overall health [181]. 201 areas in Georgia are designated as health professional shortage areas (HPSAs) for primary care, with an estimated underserved population of 1,371,292 among a total population of 2,066,893 [181]. Telemedicine programs can serve as great solutions to improving rural health due to their ability to cover more areas with fewer medical workforce.

Georgia has enacted laws that require services provided via telehealth to be reimbursed if the same service would be reimbursed when provided in person. The state has some coverage for telemedicine services in Medicaid and in private insurance. However, Georgia still has stringent standards for creating the physician-patient relationship in telemedicine encounters [182]. Physicians are less likely to offer telemedicine services to states with rural populations who are

less lucrative, more time-consuming, and have costly licensure barriers [182]. Rural areas which already have least access to health care, will remain the most underserved despite that telemedicine was developed to serve them [182].

Because of the health care disparity between Georgia's rural and urban populations, many different organizations and initiatives are starting up movements to help underserved Georgians reach better healthcare. A case study from Sustainable Rural Telehealth Innovation prioritizes what a practitioner should keep in mind when in the process of implementing telemedicine operations [183]. First, it showed the importance of seeking and following up on grants to support telehealth. Other considerations are to create independent entities with appropriate local telehealth knowledge, to tailor telehealth innovations to emerging needs and available technology options, and to facilitate participation within the rural health institution and collaboration with the local community and external partners to make the innovation sustainable [183].

7.1.2 Current State of Remote Patient Monitoring and Design Opportunities

A well-designed remote monitoring system (RMS) should be based on affordable technologies that are easily adaptable for the needs of patients. They should promote active patient engagement towards maintaining health and proper care management. Three key components are deemed important in the design of a reliable remote monitoring system. First, RMS need a secure, customizable, two-way communication module that allows patients to interact with care providers where sensitive private health care information are protected and safeguarded during transmission [184, 185]. Second, the system should be easy-to operate that integrates affordable technologies. The RMS should be able to adapt new evolving technologies readily without too much overhead. Third, it is desirable to include a movement tracking sensor for detecting sudden falls or potential health emergencies. While leading technological companies and investigators have been making

advances in RMS for patients requiring home care, most current devices lack one of these three components. Most of them are targeted towards specific chronic diseases, with a majority tailored for patients with cardiovascular disease. Table 7.1 summarizes the current state-of-the-art technology and their design platform and functionalities.

TABLE 7.1. BEST LOW-COST SENSORS, PRICE, AND CHRONIC CONDITIONS MONITORED

Sensor	Chronic Conditions	Price
Weight Gurus Digital Body Fat Scale	Diabetes, Pain (Arthritis), Heart disease, Obesity	~$40
Boso-medicus prestige BP Monitor	Diabetes, Pain (Arthritis), Heart disease, Obesity, Stroke	~$56
NeuLog™ Spirometer Sensor and NeuLog™ WiFi™ Connection Bundle	Asthma, Obesity	$153 (sensor) + $214 (WiFi connection bundle)
Accu-Chek Aviva Connect Blood Glucose Monitoring System + Test Strips	Diabetes	$30 (system) + $190 (100 test strips)
Nonin Onyx II 9560 Bluetooth Oximeter	Heart disease, Obesity, Stroke	$330
WinHealth Wireless Body Thermometer	Diabetes, Asthma, Pain (Arthritis), Heart disease, Obesity, Stroke	~$60
Rapid Response Medical Alarm with auto fall detection	Pain (Arthritis), Heart disease, Stroke	~$36/month

While some of these systems offer diverse functionalities and easy usage and are non-invasive, they are in general expensive, lack customization, and have low implementation feasibility in rural areas. An ideal system should achieve low cost, high customizability, high security, and high compliance through user-friendly design and good mobility. Furthermore, most of the existing systems lack movement tracking sensors for emergency situations. Chronic diseases such as heart disease, stroke, and arthritis can trigger sudden falls or other critical situations such as unconsciousness and inability to move. An RMS with movement and location tracking along with a panic button may facilitate immediate response/intervention by physicians/caregivers. It is possible patients may not be able to push the panic button. In this case, an auto alert module can notify the physicians/caregivers. It can be designed to connect to an emergency phone line like monitored home security systems. In this chapter, a prototypical mobile-based RMS that includes

three necessary components for effective remote patient monitoring system and chronic disease management is designed. The system can 1) measure and securely upload patient vitals through affordable and customizable sensors; 2) provide encrypted two-way communication capability between patients and physicians; 3) detect patient movements and alert healthcare providers in the event of sudden falls or emergency health situations. Our design is built on existing technologies for vital measurement sensors and a fall detection sensor. These are modified and integrated into a comprehensive system that is usable by patients with multiple conditions. Smartphone is selected as the technology platform since it has high penetration rate in the United States (estimated to reach 63.5% by 2017 [186]). The system is comprised of three modules: (1) an application for data acquisition, processing, and transmission, (2) an adaptable set of sensors for measuring vitals and reporting emergency situations, and (3) a secure communication module for remote patient-physician interactions. The user interface with the RMS is established through an application installed on the patient's smartphone.

7.1.3 Cost-Benefit Analysis

In recent years, a number of studies have pointed out return of investment (ROI) from disease management programs as a cost saving metric. ROI is a widely recognized financial tool. Its interpretation is straightforward and facilitates communication among legislators, Medicaid program officials, plan administrators, health care providers and the public. One of the advantages of using ROI is that in the analysis, it requires the inclusion of device design and manufacture and the disease program startup as well as the ongoing costs. The Center for Technology and Aging reported second-year ROIs of 2.88, 2.92, 7.24, and 1.66 respectively from four RMS studies from 2010-2012 (Table 7.2) [187]. This translates to each dollar invested in RMS yielded at least $2.88, $2.92, $7.24, and $1.66 in savings for the respective organizations. The estimated total returns per

patient in the first year for these four organizations are $1,761, $9,882, $4,388, and $2,837, respectively, as a result of reduced hospital admits and avoidance of home care visits [187]. These analyses clearly indicate that by the end of the second year, the benefits from RMS had exceeded the costs. In addition, the RMS designed in this chapter aims to reduce unnecessary costs associated with monitoring devices while ensuring quality and achieving security and compliance purposes. Therefore, it is expected that our overhead and operating costs will be lower than the analyses above, resulting in earlier and higher returns of investment.

TABLE 7.2. SUMMARY OF ROI *S*TUDIES FOR THE *F*OUR *O*RGANIZATIONS.

Organization	**Program Summary**	**ROI of RMS**	**2nd Year Total Patient Enrollme nt**
Centura Health at Home	-Diabetes, COPD, or CHF patients in Denver, CO -24/7/365 clinical call center linked with RMS utilizing Cardiocom -1.5 month average intervention length	ROI = 2.6 in Year 1, 2.88 in Year 2 Total returns/patient = $1,761	2,250
Dignity Health	-Patients with Class II, III, or IV CHF residing in Central California Coast - RMS utilizing Philips Telestation -6 month average intervention length	ROI = 0.4 in Year 1, 2.92 in Year 2 Total returns/patient = $9882	225
HealthCare Partners	-California-based patients with COPD and other chronic health conditions -Utilized low-cost interactive voice response monitoring (IVR) system -6 month average intervention length	ROI = 1.3 in Year 1, 7.24 in Year 2 Total returns/patient = 4,388	268
Sharp Healthcare	-San Diego, California-based patients with Class II or III CHF that are high utilizers of acute care and that have little or no health care insurance -Utilized Cardiocom Telescale -2.6 month average intervention length	ROI = 1.63 in Year 1, 1.66 in Year 2 Total returns/patient = 2,837	100

Due to limited generalizability, in this thesis, existing telemedicine challenges particular to the rural areas within the state of Georgia are addressed. Specifically, this chapter 1) formulates a p-median facility location to set up telemedicine sites to better serve the rural communities; 2) completes a prototypical design of a low-cost remote patient monitoring device that supports remote care management, secure communication, and monitoring of disease conditions; 3)

performs a basic cost analysis for implementation of telehealth services in rural communities in Georgia; and 4) provides recommendations for facilitating telehealth services and implementations.

7.2 MATERIAL AND METHODS

7.2.1 Optimized Location for Telemedicine Point of Distribution (POD) Facilities

7.2.1.1 PROBLEM DESCRIPTION AND FORMULATION

Currently, Georgia is approaching telemedicine is by passively waiting for rural providers to take initiative to join a bigger network in order to serve their local patients. However, a more efficient strategy is for the Georgia government to integrate a facility-location model that optimizes the system by covering all the demand for patients living in rural areas with little to no access to specialists. The main network manager (Georgia Partnership for Telehealth, for instance) might have different objectives for their system such as minimizing implementation cost, minimizing labor cost, minimizing the number of telemedicine facilities in Georgia, or maximizing the quality of care. The model described in this study focuses on the goal of minimizing the number of PODs while ensuring that the travel distance for patients from their households to a local telemedicine POD is small. Potential POD facilities include supermarkets, grocery stores, and churches. To model this problem, we follow a similar procedure described in [188]. Instead of discretizing the target region into grids, we instead use census tracts, for which the U.S. census bureau produces demographic data. Potential locations within each census tract are selected within the 6km-radius of its centroid. Let r and l be two arbitrary census tracts. Let k be the total number of census tracts in the region. We define following input parameters:

- d(r,l)_= distance between centroids r and l;
- d_{max}= maximum allowed travel distance from each household to its assigned POD location;
- c_l = the capacity of the facility at census tract l;

- p_r = demand of census tract r.

Next, we define the following decision variables:

- y_l =1 if facility site at census tract l is selected for setting up a dispensing facility, 0 otherwise;
- x_{rl} =1 if the population in census tract r is served by the facility at census tract l, 0 otherwise.

Finally, the formulation of the problem is given below:

$$\text{Minimize } \sum_{l=1}^{k} y_l \quad (1)$$

$$\text{subject to } \sum_{l=1}^{k} y_l \geq 2 \quad (2)$$

$$d(r,l)x_{rl} \leq d_{max}y_l \quad \forall\, l \in 1,\dots,k \qquad (3)$$

$$\sum_{l=1}^{k} x_{rl} = 1 \quad \forall\, r \in 1,\dots,k \quad (4)$$

$$\sum_{r=1}^{k} x_{rl}\, p_r \leq c_l y_l \forall\, l \in 1,\dots,k \qquad (5)$$

$$x_{rl}, y_l \in \{0,1\},\ Y_{ij} \in \{0,1\} \quad (7)$$

The objective function (1) minimizes the number of open POD locations. Constraint (2) ensures at least two PODs are set up. Constraint (3) sets the maximum travel capacity for each household. Constraint (4) states that each household must be served by exactly one POD facility. Constraint (5) ensures that no demand exceeds capacity at any facility. For the simplicity of the model, the following assumptions were considered:

- The cost of acquiring new telemedicine equipment is the same (constant) for every chosen facility. This implies that all the facilities will implement the same equipment and will be able to cover most basic needs of rural patients (tele-stethoscope, tele-echocardiograms, mental health resources, etc.).

- There is a budget constraint on how much money from grants is available to purchase equipment, so only a certain number of facilities can open.

7.2.1.2 CASE STUDY

This chapter investigates the 37 Georgia counties located in the Appalachian region [189]. Due to their history of economic underdevelopment, residents of these counties are particularly susceptible to chronic disease [189]. Data retrieved include: 1) US counties, census tract data and the 2010 census report from Social Explorer [190]; 2) Appalachian population health data and 3) physicians workforce data from the 2017 Health Disparities in Appalachia Report [191]; and 3) Potential PODs' names and locations (using the Google Map Places API [192]). The variables in the formulation are assigned as follows:

- Demand: There are a total of 504 census tracts within the 37 Appalachian counties. Each census tract is considered as a demand unit. The demand (p_r) for each census tract is calculated based on the available morbidity information (i.e. physically and mentally unhealthy days):

  p_r = population of census tract r * (physically unhealthy days per person per month / 30 + mentally unhealthy days per person per month / 30).

- Capacity: Here it is assumed that (1) each physician works 240 days per year and sees 18 patients per day (based on the Medical Group Management Association 2012 report), (2) each patient visits the clinic twice per year, and that (3) telemedicine practice would reduce patient visit times by 20 percent [193]. Based on this, the capacity of the facility at census tract, c_l, is calculated as:

  c_l = population of census tract r * (percentage of primary care physicians + percentage of mental health providers) * 2160 / 0.8.

- Potential POD Facilities**:** For each census tract, exactly one facility within a 6 km radius of the census tract centroid is selected as a potential POD. These facilities can be supermarkets, churches, or grocery stores.
- Distance: The centroid location of each census tract is used to calculate the pairwise distance d(r,l) between the census tracts r and l. The Haversine formula is used to calculate the pairwise distances given the latitudes and longitudes, and the unit is in kilometers.

7.2.2 Remote Patient Monitoring Device

7.2.2.1 Target Population, Market Direction, and Design Concept

Heart disease, stroke, diabetes, arthritis, and cancer are among the most common, costly, but preventable of all health problems in the U.S [194]. The target population of our RMS will be patients with one or more of these health conditions. Doctors and/or designated health coaches may utilize a remote system to monitor health progress, communicate goal planning, and promote education. With numerous sensors available today for monitoring health, uploading measurements daily to track progress overtime is a simple task. With these sensors installed, the RMS device will be especially valuable for patients with easy-to-monitor symptoms. Patients seeking preventive care may also benefit from this device. For usability and wide adoption, it is necessary to simplify disease monitoring and patient education using a simple operating system. This is especially critical for elderly patients since they may not be as familiar with smartphone technologies as the general population. Our prototype is designed to be user-friendly for all adult patients. The goal is to develop a straightforward system with clean and easy to read display menus. By using a smartphone application with straightforward displays and easy navigation, it should be easy for patients to learn to use the system, thus ensuring high adoption and compliance.

7.2.2.2 Sensors

A variety of sensors have been developed for monitoring symptoms associated with various types of chronic diseases. Peak flow meter checks levels of breathing efficiency of asthmatic patients on a daily basis. Weight scale can monitor fluctuations in weight which could be attributed to disease progression. It is also useful for patients on any weight loss regimen. Blood pressure cuffs and electrocardiogram device are available for patients with cardiovascular health issues. For diabetic patients, glucose monitor can track changes in glucose levels. PT/INR meter measures blood's anti-coagulation level – the time it takes the blood to "clot". For patients who are likely to fall due to heart attack or other health reasons, sensors that measure changes in movement pattern (e.g. movement acceleration, changes in height and orientation to the horizontal position) can detect sudden falls. Sensor readings can be collected and input into an aggregator system manually, but the goal of our prototype is to automatically retrieve sensor measurements via wireless internet or Bluetooth. The list of sensors typically used by patients with the most common chronic diseases is shown in Table 7.3.

TABLE 7.3. LIST OF GENERAL SENSORS AND THE CHRONIC CONDITIONS MONITORED.

Sensor	Chronic Conditions
Scale	Diabetes, Pain (Arthritis), Heart disease, Obesity
Blood Pressure Monitor	Diabetes, Pain (Arthritis), Heart disease, Obesity, Stroke
Spirometer	Asthma, Obesity
Glucose Monitor	Diabetes
Heart Monitor	Heart disease, Obesity, Stroke
Body Temperature Monitor	Diabetes, Asthma, Pain (Arthritis), Heart disease, Obesity, Stroke
Movement Tracking Sensor	Pain (Arthritis), Heart disease, Stroke

Many sensors have been incorporated into wearable systems and have become increasingly unobtrusive [195]. Movement sensors today are inexpensive, small, and require little power [195]. They will be incorporated into our system for monitoring emergency situations such as sudden falls. However other sensors can still be too costly and difficult to use, and this reduces the

feasibility for deployment in rural areas and adoption/compliance level. Sensors listed in Table 24 are selected primarily based on financial considerations while ensuring robust quality and easy operation. In order to achieve lowest cost in deployment of our RMS, the sensor system is personalized for patients according to their individual conditions.

7.2.2.3 COMMUNICATION MODULE

The communication module is set up to allow the patient to interact with medical staff remotely. By using the webcam on the smartphone, the patient can hold a video conference with the doctor rather than traveling to the hospital to ask a simple question or discuss treatment progress face-to-face. If the patient's vital signs raise a red flag in the system, a nurse or physician assistant can also contact the patient to review medication and goals. One major technical challenge in patient-physician communication is that rural broadband connectivity costs for services are higher than metropolitan rates. However, the communication technologies themselves are much easier to use than anticipated. The expansion of 4G and LTE networks in the United States helps solve connectivity issues for higher quality video. Unfortunately, due to security concerns, constant or even frequent open two-way communication pathways are difficult with current security protocols and technology. Hackers and others with malicious intent could easily gain access to sensitive healthcare information if such channels were open. Nevertheless, FaceTime, Skype, and Adobe Connect had implemented encryption and other security measures. These encryptions (and their continued advances) are indispensable for the usage of RMS due to the sensitivity of the transmitted data. Our RMS incorporates Skype's AES-style encryption. Skype-to-Skype calling and video calling products ensure that Protected Health Information (PHI) are encrypted and is free to use and easily installed on smart mobile devices [196].

7.2.2.4 DATA ACQUISITION AND PROCESSING MODULE

The TeleCARE system [197] has proven successful in the remote monitoring process for treatment of patients with cardiovascular diseases. Our design integrates two of its components into our system – the data acquisition module, which is responsible for collecting data from patient's medical devices and sending them for further processing, and the data processing module, which is used for receiving measured data from patients and providing analysis and visualization according to rules previously individually defined by physicians. In addition, we integrate the TeleCARE mobile application, which is designed in a modular manner that allows for extensions of devices to connect to the core application itself and provides measurement values. Currently, it is implemented on Android OS, offering functionalities such as simple implementation and pairing, geo-localization, measurement visualization and data processing. It will also be implemented on the iOS platform, and its security measures will be enhanced when transmitting data between patient's devices and physicians.

7.2.2.5 SMARTPHONE APPLICATION

To establish a user-friendly and highly mobile system, we implement a smartphone application that provides an interface to accessing all these technologies. The application will require fingerprint/password access to protect patient identity. It will be implemented on the most prevalent Android and iOS platforms.

7.3 RESULTS

7.3.1 Solution to Optimized Telemedicine POD Problem

The AMPL [198] system is used to build the optimization model using the formulation in the materials and methods section, and CPLEX for AMPL 12.2 is used to solve the model. Python scripts are developed to parse each dataset and calculate the pairwise Haversine distances. The generated data are used to prepare the AMPL data file.

When d_{max} is set to10 km, the optimal solution for the objective function is 126. This represent that the minimum of 126 PODs are required to serve the 504 census tracts. When d_{max} is set to 20 km, the number of PODs required reduces to only 35.

7.3.2 Prototypic Design of Remote Patient Monitoring Device and Return of Investment Analysis

7.3.2.1 Prototype Architecture of RMS

Based on the existing sensor technologies, communication tools, and data management applications, we integrate and design our system. Figure 7.1 shows the prototype architecture of our RMS. Customized sensors for individual patients take vital measurements and upload them to both the smartphone application and the secure cloud storage. Patients manage their chronic conditions through the smartphone application. Physicians download patient vitals securely through the cloud storage. Two-way communication between physicians and patients is achieved via Skype's video conferencing technologies.

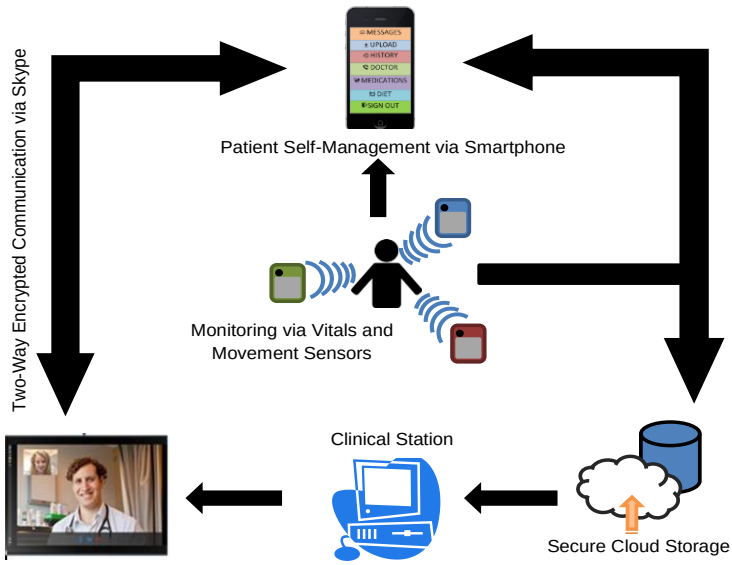


Figure 7.1. Main modules of the remote patient monitoring system.

The purpose of the design is multi-fold. It is intended to increase the patient's independence and proactivity in his/her own health care management, while facilitating communication when needed between patient and physician for intervention. The intelligent communication pathway through broadband is used to facilitate communication between the patient, healthcare providers, and other third-party providers. These third-party providers include family and support groups, employer wellness programs, and pharmaceutical suppliers. In the backend, nurses, doctors, and other care managers can work collaboratively to give the patient appropriate feedback and support.

7.3.2.2 SMARTPHONE MANAGEMENT APPLICATION

The mobile application is implemented and deployed on the most popular Android and iOS platforms found on modern smartphones. At home, the patient logs on securely with unique bio ID or username and password to the health self-management application to collect vital sign data through sensor devices and manage treatment process. Through the application, the patient can receive a personalized healthcare plan, which includes messages from healthcare providers and information regarding treatment goals, medications, diet, and physical activities. The patient can also access vital history charts and upload additional vital readings to the cloud database. Figure *7.2* shows the design of the main interface of the smartphone management application.

Figure 7.2. Main user interface of the smartphone management application.

Each of the menu functions are explained below.

- Messages: Users can compose or view messages. The messages sent from the doctor will be text-based, and the outgoing messages can be text, audio, or video-based.
- Upload: Although sensors can be programmed to automatically collect readings and upload data to the cloud database, this option allows user to take and upload the vital measurements whenever they want.
- History: History containing charts of the patient's sensor readings over time can be visualized.
- Doctor: It contains emergency contact information for the hospital. The user can start a video conference with their doctor or take photos for quick remote diagnosis, minimizing the need for hospital visits.
- Medications: Users can view medications prescribed and when to take them. Doctors can create automatic reminders for patients to take medications on time.
- Diet: Users can view daily diet plans designed for them.

- Sign Out: Signs out and exits the management program. This prevents sensitive medication from leaking when Smartphones are unattended. The application will automatically sign out after being idle for a while.

7.3.2.3 ROI ESTIMATION FOR CARE COORDINATION INSTITUTE'S PATIENT COHORT

To evaluate the feasibility of implementing the RMS designed in this study, an ROI estimation on 2.7 million patients from the Care Coordination Institute is performed. Among these patients, almost 1 million have more than one chronic condition, and more than 0.1 million have more than three chronic conditions. Using the technology operation, management, and personnel costs listed in Table 7.4, year one ROIs of 0.65, 0.03, 0.19, 0.13, 0.11, 0.3, and 0.57 are calculated respectively for patients with incrementing number of chronic conditions (Table 7.5). This analysis shows that each dollar invested in RMS for an individual without any chronic conditions is projected to yield at least $0.65 in savings ($1.65 net return) in year 1. As anticipated, the savings is also significant among patients with chronic diseases. As more patients start to use the RMS service and the technology operations become more automatic, we expect ROIs for later years to continue to increase.

TABLE 7.4. TECHNOLOGY OPERATION, MANAGEMENT, AND PERSONNEL COSTS PER PATIENT.

Software System Purchase Cost	Hardware System Purchase Cost (Per Chronic Condition)	System Management Cost	Personnel Costs (Consulting, Survey, Patient Education)	Mean Out-of- Pocket Spending
$100	$200 base price with additional charge on certain components	Estimates based on patient population	$150	Based on [174, 175] and the number of chronic conditions: no condition: $343; one: $795; two: $1,534; three:1,760; four: 1,814; five: 2,314; six: 3,815; > seven:4,217

TABLE 7.5. ROI ANALYSIS BY NUMBER OF CHRONIC CONDITIONS.

Number of Chronic Conditions	Mean Out-of-Pocket Spending [175]	Year 1 Total Savings	Year 1 ROI	Year 2 ROI
0	$343	$407,588,600	0.65	0.82
1	$795	$7,475,786	0.03	0.82
2	$1,134	$37,867,817	0.19	0.75
3	$1,760	$11,414,186	0.13	0.66
4	$1,814	$1,970,436	0.11	0.63
5	$2,314	$653,447	0.3	0.71
6	$3,815	$88,790	0.57	0.82

7.4 DISCUSSION

In this study, we identified the key benefits and challenges in the field of telemedicine and systematically explored its state-of-art in Georgia. Based on existing knowledge and technologies, we provided a practical system for expanding telehealth services in the rural communities of Georgia by optimizing potential service locations.

While telehealth is growing and expanding in a number of positive directions, observations of the current system provide many suggestions on what measures can be taken to improve telehealth implementation. Hospitals need to implement telehealth through appropriate key performance indicators and allocation of necessary resources. Hospitals also need to incorporate telehealth as part of their core business because a solid business plan with a positive ROI is critical to integrate a telehealth program even though impact on access to care and quality of care is important. Providers should be able to more easily offer care to patients in multiple states. Telehealth services implemented with well-defined patient inclusion criteria and protocol would be cost-effective. An increase in safety and quality concerns has also proven to reinvigorate and jumpstart programs which can create a greater focus on being cost-effective. Continued grant funding can help appease a lack of a physician buy-in and consequentially expand the number of access points and the number of participating healthcare providers within the telemedicine network.

Based on the current challenges associated with funding, primary care, and healthcare services in Georgia, the following recommendations are identified to alleviate the situation:

(1) Secure additional sources of funding. Between 2010 and 2030, Georgia's population is projected to grow by an additional 4.6 million people [199]. In 2011 almost 40% of Georgia hospitals lost money, and 55% Georgia rural hospitals had negative total margins [181]. It is therefore extremely important to individual and providers as well as big networks to develop a sustainable business model to implement telemedicine as an effective cost-saving strategy by reaching out to telemedicine grants.

(2) Expand network to address its shortage of primary care providers. One method is to recruit and survey additional physicians. The Georgia Volunteer Health Care Program (GVHCP) provides sovereign immunity protection to licensed healthcare professionals who volunteer to treat uninsured individuals at or below 200% of the federal poverty level [200]. Another solution is to increase the flexibility in state licensing regulations so that doctors from states with a higher physician-to-patient ratio could provide telemedicine care to patients in states like Georgia.

(3) Increase available services. Finally, Georgia's specific needs demands an expansion in the areas of tele-dentistry, high risk OB/Centering pregnancy, division of family & children services (DFACS), and school-based telemedicine programs.

In this chapter, a prototypic design of a personalized remote patient monitoring system is also presented. The design focuses on patients with one or more chronic diseases. Initially, the diseases targeted include asthma, diabetes, chronic pain/arthritis, cardiovascular disease, obesity, and stroke. Having assessed available monitoring sensors, data management and communication technologies, a new system is composed with a smart-phone based management application. RMS addresses growing needs of device and disease compatibility, secure two-way communication, as

well as affordability and adoption/compliance. In addition to increasing patient involvement in their own healthcare, there is a need to increase patient and provider communication. Interaction between doctors and their patients can be facilitated at a lower cost via a remote monitoring system than attending a face-to-face consultation. Patients can receive timely medical advice using RMS, avoiding the unnecessary wait for appointment that may take several days or weeks. While the potential benefits of RMS may outweigh the costs, there are some limitations to keep in mind. There is a high initial one-time purchase cost and the subsequent maintenance costs as well as the cost for training the hospital staff. Even though the overhead cost may be high, it is estimated that the system would pay for itself within two years. Doctors will be able to keep track of patients' progress without lengthy and at times delayed consultations. By implementing this system, the cost of healthcare, particularly in the area of chronic diseases, will significantly drop. In the 2015 Institute of Medicine "Transforming Health Care Scheduling and Care, Getting to Now" report [201], it was stated that various technologies are emerging with strong potential to improve real-time access to care, with the promise of totally new ways of scheduling and delivering care and gathering information on its utility. Use of digital and social media, telemedicine and telehealth, remote monitoring, and related evolving technologies are also well suited for deployment in health care practices. Still, their uptake has been relatively limited to date, for such reasons as unfamiliarity, system mismatch, and absence of reimbursement. Underlying the geographic and physical barriers to access is the reliance of the U.S. health care system on the office visit as the default model of care. Our RMS can thereby offer timely alternative care which can be greatly beneficial to chronic disease patients. This is particularly true when RMS can encourage proactive engagement of the patients in their own healthcare management. The ability to consult remotely in a timely manner allows them to maintain a steady health condition between office visits. For

the rural population, RMS will relieve the lack-of-access and expertise burdens within the region. Overall, there is strong evidence that a remote monitoring system will benefit those with chronic diseases. It allows a reduction in personal healthcare costs while increasing overall wellness. Doctors would view the system as a useful tool for reducing hospital readmission rates while keeping lines of patient communication open. Though our approach is tailored to meet the needs of today's market, we expect the Smartphone-based system will continue to appeal to the growing, aging population as an entertaining and motivating way for patients to take charge of their personal heath.

CHAPTER VIII

CONCLUSION AND FUTURE WORK

In this dissertation, a systematic analytic pipeline is developed to address the challenges associated with big data in healthcare. This pipeline is composed of technologies and innovations related to information extraction, data pre-processing, terminology standardization, longitudinal data mining, feature selection and supervised learning, and telehealth. Each of these technologies are developed fill some of the most important gaps in the current healthcare industry.

In chapter 3, a comprehensive information extraction is developed for Electronic Health Records. The pipeline organizes data into a structured, machine-readable format which can be effectively applied in clinical research studies to generate practical and unbiased insights from a holistic perspective.

Chapter 4 builds an automated mapping process to map unstructured clinical texts to standardized SNOMED-CT concepts. The process establishes interoperability among the heterogeneous data systems from multiple health care sites and providers. It significantly reduces data dimension and redundancy, thereby addressing the issue of "curse of dimensionality" associated with big data. It also has the potential to facilitate the reimbursement process, improve clinical decision support, and promote the exchange of information among multiple sites.

Chapter 5 explores the vastly under-utilized longitudinal health data. In this chapter, various distance metrics and clustering methods for irregular time series data are investigated. These methods address some of the most challenging aspects associated with clinical laboratory measurement data—sparse, unevenly-spaced, and unequal in length. Furthermore, an approach to cluster irregular multivariate time series data is developed. Future works remains in the search of

more robust and systematic methods for evaluating the quality of time series clusters. Given the complexity of irregular MTS and the difficulty involved in labelling clusters, it is necessary to combine effective visualization techniques with quantitative measures to achieve this task. Mining these multi-dimensional longitudinal health data is conducive to understanding patient conditions from an evolving perspective. It opens new doors to improving clinical outcome assessment, prognostic tools, and personalized care.

In chapter 6, clinical outcome prediction models are built with various state-of-art supervised learning algorithms. Structured and unstructured data are extracted through the information extraction pipeline built in chapter 3. Input features are obtained with the automatic concept mapping system developed in chapter 4, and treatment outcome is defined by time series clustering approaches developed in chapter 5. Feature selection algorithms are leveraged to identify small number of critical factors that affect treatment outcome. Classification rules are established via machine learning models using these critical factors. These rules facilitate the design of practical evidence-based treatment guidelines and optimization of site performance through best practice dissemination and knowledge transfer.

Chapter 7 identifies and addresses current challenges in health disparity and leverages telehealth and remote patient monitoring as practical intervention methods. Specifically, an optimization model to set up telehealth point-of-service locations is formulated and solved, and a remote patient monitoring system prototype is designed. Using a return of investment cost-benefit analysis on a cohort of 2.7 million patients, the feasibility of the system is evaluated. An important future task will involve finding the critical value of maximum allowed travel distance that can better balance the tradeoff between the number of PODs and the distance. By addressing the challenges in the area of telemedicine and expanding its influence, healthcare industries will be

able to improve efficiency and timeliness of patient-centered care, serve more patients in need, improve rural health by increasing access to care, reducing unnecessary face-to-face visits, and reducing costs.

The analytic pipeline developed in this thesis has been applied to the 2.7 million patient cohort in the CCI-Health database and the EPIC EHR system. It can be adapted to generic EHR systems and clinical big data. Future advances will involve 1) exploring more efficient, comprehensive, and accurate information extraction approaches, 2) identifying integrating more sources of clinical vocabularies into standardized terminology systems, 3) improving the efficiency of time series clustering algorithms and developing more rigorous evaluation metrics for clinical time series clustering results, 4) understanding the role of various laboratory measurements as biomarkers of risk and recovery progress, 5) advancing feature selection and supervised learning methods to improve model performance while addressing issues with clinical datasets such as imbalanced classes, 6) monitoring health status and performing rapid analytics based on existing data to identify the first sign of health risk, 7) implementing practical clinical intervention strategies based on learned knowledge from data mining, and 8) promoting patient-centered pre-disease and behavior intervention. Finally, the velocity of big data makes it increasingly important to develop advanced analytical methods that can support real-time decision making.

REFERENCE

1. Gillum, R.F., *From papyrus to the electronic tablet: a brief history of the clinical medical record with lessons for the digital age.* The American journal of medicine, 2013. **126**(10): p. 853-857.
2. Gunter, T.D. and N.P. Terry, *The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions.* Journal of medical Internet research, 2005. **7**(1): p. e3.
3. Henry, J., et al., *Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2015.* ONC data brief, 2016. **35**: p. 1-9.
4. Kumar, K.M., S. Tejasree, and S. Swarnalatha. *Effective implementation of data segregation & extraction using big data in E-health insurance as a service*. in *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2016. IEEE.
5. Health, U.D.o. and H. Services, *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* US Department of Health and Human Services, Washington, DC) Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed September, 2012. **26**: p. 2018.
6. Lee, E.K., et al. *Machine learning: Multi-site evidence-based best practice discovery*. in *International Workshop on Machine Learning, Optimization, and Big Data*. 2016. Springer.
7. Birman-Deych, E., et al., *Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors.* Medical care, 2005: p. 480-485.
8. Kern, E.F., et al., *Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes.* Health services research, 2006. **41**(2): p. 564-580.
9. Li, L., et al. *Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study*. in *AMIA Annual Symposium Proceedings*. 2008. American Medical Informatics Association.
10. Savova, G.K., et al. *Discovering peripheral arterial disease cases from radiology notes using natural language processing*. in *AMIA Annual Symposium Proceedings*. 2010. American Medical Informatics Association.
11. Long, W. *Extracting diagnoses from discharge summaries*. in *AMIA annual symposium proceedings*. 2005. American Medical Informatics Association.
12. Turchin, A., et al., *Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes.* Journal of the American Medical Informatics Association, 2006. **13**(6): p. 691-695.
13. Friedlin, J. and C.J. McDonald. *A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports*. in *AMIA annual symposium proceedings*. 2006. American Medical Informatics Association.
14. Baud, R.H., et al. *Morpho-semantic parsing of medical expressions*. in *Proceedings of the AMIA Symposium*. 1998. American Medical Informatics Association.
15. Mamlin, B.W., D.T. Heinze, and C.J. McDonald. *Automated extraction and normalization of findings from cancer-related free-text radiology reports*. in *AMIA Annual Symposium Proceedings*. 2003. American Medical Informatics Association.
16. Friedman, C., et al., *A general natural-language text processor for clinical radiology.* Journal of the American Medical Informatics Association, 1994. **1**(2): p. 161-174.
17. Jain, N.L. and C. Friedman. *Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports*. in *Proceedings of the AMIA Annual Fall Symposium*. 1997. American Medical Informatics Association.
18. Bashyam, V. and R.K. Taira. *Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA*. in *AMIA Annual Symposium Proceedings*. 2005. American Medical Informatics Association.

19. Taira, R.K. and S.G. Soderland. *A statistical natural language processor for medical reports*. in *Proceedings of the AMIA Symposium*. 1999. American Medical Informatics Association.
20. Settles, B. *Biomedical named entity recognition using conditional random fields and rich feature sets*. in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*. 2004.
21. Rindflesch, T.C., et al., *EDGAR: extraction of drugs, genes and relations from the biomedical literature*, in *Biocomputing 2000*. 1999, World Scientific. p. 517-528.
22. Friedman, C., et al. *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*. in *ISMB (supplement of bioinformatics)*. 2001.
23. GuoDong, Z. and S. Jian. *Exploring deep knowledge resources in biomedical name recognition*. in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. 2004. Association for Computational Linguistics.
24. Kazama, J.i., et al. *Tuning support vector machines for biomedical named entity recognition*. in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*. 2002. Association for Computational Linguistics.
25. Mitsumori, T., et al., *Gene/protein name recognition based on support vector machine using dictionary as features.* BMC bioinformatics, 2005. **6**(1): p. S8.
26. Boag, W., et al., *CliNER 2.0: Accessible and Accurate Clinical Concept Extraction.* arXiv preprint arXiv:1803.02245, 2018.
27. Chalapathy, R., E.Z. Borzeshi, and M. Piccardi, *Bidirectional LSTM-CRF for clinical concept extraction.* arXiv preprint arXiv:1611.08373, 2016.
28. Mutalik, P.G., A. Deshpande, and P.M. Nadkarni, *Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS.* Journal of the American Medical Informatics Association, 2001. **8**(6): p. 598-609.
29. Chapman, W.W., et al., *A simple algorithm for identifying negated findings and diseases in discharge summaries.* Journal of biomedical informatics, 2001. **34**(5): p. 301-310.
30. Gindl, S., K. Kaiser, and S. Miksch, *Syntactical negation detection in clinical practice guidelines.* Studies in health technology and informatics, 2008. **136**: p. 187.
31. Elkin, P.L., et al., *A controlled trial of automated classification of negation from clinical notes.* BMC medical informatics and decision making, 2005. **5**(1): p. 13.
32. De Bruijn, B., et al., *Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010.* Journal of the American Medical Informatics Association, 2011. **18**(5): p. 557-562.
33. Díaz, N.P.C., et al., *A machine-learning approach to negation and speculation detection in clinical texts.* Journal of the American society for information science and technology, 2012. **63**(7): p. 1398-1410.
34. Goldin, I. and W.W. Chapman. *Learning to detect negation with 'not' in medical texts*. in *Proc Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*. 2003.
35. Zikopoulos, P. and C. Eaton, *Understanding big data: Analytics for enterprise class hadoop and streaming data*. 2011: McGraw-Hill Osborne Media.
36. Sweeney, L. *Replacing personally-identifying information in medical records, the Scrub system*. in *Proceedings of the AMIA annual fall symposium*. 1996. American Medical Informatics Association.
37. Ruch, P., et al. *Medical document anonymization with a semantic lexicon*. in *Proceedings of the AMIA Symposium*. 2000. American Medical Informatics Association.
38. Taira, R.K., A.A. Bui, and H. Kangarloo. *Identification of patient name references within medical documents using semantic selectional restrictions*. in *Proceedings of the AMIA Symposium*. 2002. American Medical Informatics Association.
39. Sibanda, T. and O. Uzuner. *Role of local context in automatic deidentification of ungrammatical, fragmented text*. 2006. Association for Computational Linguistics.
40. Wellner, B., et al., *Rapidly retargetable approaches to de-identification in medical records.* Journal of the American Medical Informatics Association, 2007. **14**(5): p. 564-573 %@ 1527-974X.

41. Szarvas, G., R. Farkas, and R. Busa-Fekete, *State-of-the-art anonymization of medical records using an iterative machine learning framework.* Journal of the American Medical Informatics Association, 2007. **14**(5): p. 574-580 %@ 1527-974X.
42. Rosenbloom, S.T., et al., *Interface terminologies: facilitating direct entry of clinical data into electronic health record systems.* Journal of the American medical informatics association, 2006. **13**(3): p. 277-288.
43. Osornio, A.L., et al., *Creation of a local interface terminology to SNOMED CT.* Studies in health technology and informatics, 2007. **129**(1): p. 765.
44. Wang, S.J., et al., *Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool.* International journal of medical informatics, 2003. **72**(1-3): p. 17-28.
45. Green, J.M., et al., *Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT® post-coordination.* Journal of the American Medical Informatics Association, 2006. **13**(3): p. 321-333.
46. Herbert, I., *CLICSIG report: issues around compositional terminologies, SNOMED-CT in particular.* Informatics in primary care, 2007. **15**(3): p. 193-197.
47. Nachimuthu, S. and L.M. Lau. *Practical issues in using SNOMED CT as a reference terminology*. in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. 2007. IOS Press.
48. Donnelly, K., *SNOMED-CT: The advanced terminology and coding system for eHealth.* Studies in health technology and informatics, 2006. **121**: p. 279.
49. Bodenreider, O., *The unified medical language system (UMLS): integrating biomedical terminology.* Nucleic acids research, 2004. **32**(suppl_1): p. D267-D270.
50. Aronson, A.R. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. in *Proceedings of the AMIA Symposium*. 2001. American Medical Informatics Association.
51. Schuyler, P.L., et al., *The UMLS Metathesaurus: representing different views of biomedical concepts.* Bulletin of the Medical Library Association, 1993. **81**(2): p. 217.
52. (US), B.M.N.L.o.M. *UMLS® Reference Manual*. 2009 [cited 2019 04-30]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK9684/.
53. Forrey, A.W., et al., *Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results.* Clinical Chemistry, 1996. **42**(1): p. 81-90.
54. Khan, A.N., et al., *Standardizing laboratory data by mapping to LOINC.* Journal of the American Medical Informatics Association, 2006. **13**(3): p. 353-355.
55. Vreeman, D.J. and C.J. McDonald. *Automated mapping of local radiology terms to LOINC*. in *AMIA Annual Symposium proceedings*. 2005. American Medical Informatics Association.
56. Zunner, C., et al., *Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V. 5: a semi-automated approach.* Journal of the American Medical Informatics Association, 2012. **20**(2): p. 293-297.
57. Liu, S., et al., *RxNorm: prescription for electronic drug information exchange.* IT professional, 2005. **7**(5): p. 17-23.
58. Nelson, S.J., et al., *Normalized names for clinical drugs: RxNorm at 6 years.* Journal of the American Medical Informatics Association, 2011. **18**(4): p. 441-448.
59. Bodenreider, O. *Issues in mapping LOINC laboratory tests to SNOMED CT*. in *AMIA Annual Symposium Proceedings*. 2008. American Medical Informatics Association.
60. Hernandez, P., et al. *Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse*. in *AMIA Annual Symposium Proceedings*. 2009. American Medical Informatics Association.
61. Saitwal, H., et al., *Cross-terminology mapping challenges: a demonstration using medication terminological systems.* Journal of biomedical informatics, 2012. **45**(4): p. 613-625.

62. Trott, P., *International classification of diseases for oncology.* Journal of clinical pathology, 1977. **30**(8): p. 782.
63. Carlo, L., H.S. Chase, and C. Weng. *Aligning structured and unstructured medical problems using umls*. in *AMIA Annual Symposium Proceedings*. 2010. American Medical Informatics Association.
64. Patel, C.O. and J.J. Cimino, *Using semantic and structural properties of the unified medical language system to discover potential terminological relationships.* Journal of the American Medical Informatics Association, 2009. **16**(3): p. 346-353.
65. Miaskowski, C., et al. *Subgroups of patients with cancer with different symptom experiences and quality-of-life outcomes: a cluster analysis*. 2006.
66. Wells, B.J., et al., *Strategies for handling missing data in electronic health record derived data.* Egems, 2013. **1**(3).
67. Marlin, B.M., et al. *Unsupervised pattern discovery in electronic health care data using probabilistic clustering models*. 2012. ACM.
68. Lee, C.F., J.C. Lee, and A.C. Lee, *Statistics for business and financial economics*. Vol. 1 %@ 9810234856. 2000: Springer.
69. Kreindler, D.M. and C.J. Lumsden, *The effects of the irregular sample and missing data in time series analysis*, in *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data*. 2016, CRC Press. p. 149-172.
70. Carlstein, E., *Resampling techniques for stationary time-series: some recent developments.* IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS, 1992. **45**: p. 75-75 %@ 0940-6573.
71. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108 %@ 0035-9254.
72. Agrawal, R., C. Faloutsos, and A. Swami. *Efficient similarity search in sequence databases*. 1993. Springer.
73. Chan, K.-P. and A.W.-C. Fu. *Efficient time series matching by wavelets*. 1999. IEEE.
74. Yi, B.-K. and C. Faloutsos. *Fast time sequence indexing for arbitrary Lp norms*. 2000. Citeseer.
75. Latecki, L.J., et al. *Elastic partial matching of time series*. 2005. Springer.
76. Marteau, P.-F., *Time warp edit distance with stiffness adjustment for time series matching.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009. **31**(2): p. 306-318 %@ 0162-8828.
77. Vlachos, M., G. Kollios, and D. Gunopulos. *Discovering similar multidimensional trajectories*. in *Proceedings 18th international conference on data engineering*. 2002. IEEE.
78. Sakoe, H. *Dynamic-programming approach to continuous speech recognition*. in *1971 Proc. the International Congress of Acoustics, Budapest*. 1971.
79. Smyth, P. *Clustering sequences with hidden Markov models*. in *Advances in neural information processing systems*. 1997.
80. Kalpakis, K., D. Gada, and V. Puttagunta. *Distance measures for effective clustering of ARIMA time-series*. in *Proceedings 2001 IEEE international conference on data mining*. 2001. IEEE.
81. Das, G., D. Gunopulos, and H. Mannila. *Finding similar time series*. in *European Symposium on Principles of Data Mining and Knowledge Discovery*. 1997. Springer.
82. Huhtala, Y., J. Karkkainen, and H.T. Toivonen. *Mining for similarities in aligned time series using wavelets*. in *Data Mining and Knowledge Discovery: Theory, Tools, and Technology*. 1999. International Society for Optics and Photonics.
83. Chen, L. and R. Ng. *On the marriage of lp-norms and edit distance*. in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. 2004. VLDB Endowment.
84. Cuturi, M., et al. *A kernel for time series based on global alignments*. in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. 2007. IEEE.
85. Brockwell, P.J., R.A. Davis, and M.V. Calder, *Introduction to time series and forecasting.* Vol. 2. 2002: Springer.
86. Singhal, A. and D.E. Seborg, *Clustering multivariate time-series data.* Journal of Chemometrics: A Journal of the Chemometrics Society, 2005. **19**(8): p. 427-438.

87. Liao, T.W., *A clustering procedure for exploratory mining of vector time series.* Pattern Recognition, 2007. **40**(9): p. 2550-2562.
88. Košmelj, K. and V. Batagelj, *Cross-sectional approach for clustering time varying data.* Journal of Classification, 1990. **7**(1): p. 99-109.
89. Ramoni, M., P. Sebastiani, and P. Cohen, *Bayesian clustering by dynamics.* Machine learning, 2002. **47**(1): p. 91-121.
90. Amarasingham, R., et al., *Implementing electronic health care predictive analytics: considerations and challenges.* Health Affairs, 2014. **33**(7): p. 1148-1154.
91. Taylor, R.A., et al., *Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach.* Academic emergency medicine, 2016. **23**(3): p. 269-278.
92. Wu, J., J. Roy, and W.F. Stewart, *Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches.* Medical care, 2010: p. S106-S113.
93. Gultepe, E., J.P. Green, and H. Nguyen, *From vital signs to clinical outcomes for*. 2013.
94. Kawaler, E., et al. *Learning to predict post-hospitalization VTE risk from EHR data*. in *AMIA annual symposium proceedings*. 2012. American Medical Informatics Association.
95. Panahiazar, M., et al., *Using EHRs and machine learning for heart failure survival analysis.* Studies in health technology and informatics, 2015. **216**: p. 40.
96. Zhai, H., et al., *Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children.* Resuscitation, 2014. **85**(8): p. 1065-1071.
97. Asadi, H., et al., *Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy.* PloS one, 2014. **9**(2): p. e88225.
98. Cruz, J.A. and D.S. Wishart, *Applications of machine learning in cancer prediction and prognosis.* Cancer informatics, 2006. **2**: p. 117693510600200030.
99. Khalilia, M., S. Chakraborty, and M. Popescu, *Predicting disease risks from highly imbalanced data using random forest.* BMC medical informatics and decision making, 2011. **11**(1): p. 51.
100. Månsson, K.N., et al., *Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning.* Translational psychiatry, 2015. **5**(3): p. e530.
101. Cohen, G., et al., *Learning from imbalanced data in surveillance of nosocomial infection.* Artificial intelligence in medicine, 2006. **37**(1): p. 7-18.
102. Passos, I.C., et al., *Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach.* Journal of affective disorders, 2016. **193**: p. 109-116.
103. Wall, D.P., et al., *Use of machine learning to shorten observation-based screening and diagnosis of autism.* Translational psychiatry, 2012. **2**(4): p. e100.
104. Lee, T.-F., et al., *Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer.* PloS one, 2014. **9**(2): p. e89700.
105. Lee, T.-F., et al., *LASSO NTCP predictors for the incidence of xerostomia in patients with head and neck squamous cell carcinoma and nasopharyngeal carcinoma.* Scientific reports, 2014. **4**: p. 6217.
106. Sun, Y., J. Yao, and S. Goodison. *Feature selection for nonlinear regression and its application to cancer research*. in *Proceedings of the 2015 SIAM International Conference on Data Mining*. 2015. SIAM.
107. Lee, E.K., et al. *A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department*. in *AMIA Annual Symposium Proceedings*. 2012. American Medical Informatics Association.
108. Field, M.J., *Telemedicine: A guide to assessing telecommunications for health care*. 1996: National Academies Press.

109. Hu, P.J.-H., P.Y. Chau, and O.R.L. Sheng, *Adoption of telemedicine technology by health care organizations: an exploratory study*. Journal of organizational computing and electronic commerce, 2002. **12**(3): p. 197-221.
110. Sable, C.A., et al., *Impact of telemedicine on the practice of pediatric cardiology in community hospitals.* Pediatrics, 2002. **109**(1): p. E3.
111. Ditchburn, J.L. and A. Marshall, *Renal telemedicine through video-as-a-service delivered to patients on home dialysis: A qualitative study on the renal care team members' experience.* Journal of renal care, 2017. **43**(3): p. 175-182.
112. Selkirk, S.M., et al., *Delivering tertiary centre specialty care to ALS patients via telemedicine: a retrospective cohort analysis.* Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 2017. **18**(5-6): p. 324-332.
113. Hersh, W., et al., *Telemedicine for the Medicare population. Evidence report/technology assessment no. 24. AHRQ publication no. 01-E012*, in *Rockville, MD: Agency for Healthcare Research and Quality;*. 2001.
114. Hersh, W.R., et al., *Clinical outcomes resulting from telemedicine interventions: a systematic review.* BMC medical informatics and decision making, 2001. **1**(1): p. 5.
115. Lee, E.K., et al. *Designing a low-cost adaptable and personalized remote patient monitoring system*. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017. IEEE.
116. Scott Kruse, C., et al., *Evaluating barriers to adopting telemedicine worldwide: A systematic review.* Journal of telemedicine and telecare, 2018. **24**(1): p. 4-12.
117. Olanrewaju, R., et al., *ICT in Telemedicine: conquering privacy and security issues in health care services.* Electronic Journal of Computer Science and Information Technology, 2013. **4**(1): p. 19-24.
118. Organization, W.H., *Telemedicine: opportunities and developments in member states. Report on the second global survey on eHealth*. 2010: World Health Organization.
119. Svorny, S., *Liberating Telemedicine: Options to Eliminate the State-Licensing Roadblock.* Cato Institute. Nov, 2017. **15**.
120. Thomas, L. and G. Capistrant, *State telemedicine gaps analysis: Physician practice standards & licensure.* American Telemedicine Association, 2016.
121. Chang, S.L., J.M. Pines, and J.H. Thorpe, *How the European Union Is Embracing Cross-border Telemedicine and what the US State Medical Boards Can Learn From It.* 2018.
122. Yamamoto, D.H., *Assessment of the feasibility and cost of replacing in-person care with acute care telehealth services.* Alliance for Connected Care, December, 2014.
123. Sanders, J.H. and R.L. Bashshur, *Challenges to the implementation of telemedicine.* Telemedicine Journal, 1995. **1**(2): p. 115-123.
124. Neufeld, J.D., C.R. Doarn, and R. Aly, *State policies influence Medicare telemedicine utilization.* Telemedicine and e-Health, 2016. **22**(1): p. 70-74.
125. Weisgrau, S., *Issues in rural health: Access, hospitals, and reform.* Health care financing review, 1995. **17**(1): p. 1.
126. Logan, A.G., et al., *Mobile phone–based remote patient monitoring system for management of hypertension in diabetic patients.* American journal of hypertension, 2007. **20**(9): p. 942-948.
127. Wu, X., et al., *Data mining with big data.* IEEE transactions on knowledge and data engineering, 2014. **26**(1): p. 97-107.
128. Waitman, L.R., et al. *Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement*. in *AMIA Annual Symposium Proceedings*. 2011. American Medical Informatics Association.
129. Honnibal, M. and M. Johnson. *An improved non-monotonic transition system for dependency parsing*. in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015.
130. Mihalcea, R. and P. Tarau. *Textrank: Bringing order into text*. in *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

131. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 2018. **68**(6): p. 394-424.
132. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019.* CA: a cancer journal for clinicians, 2019. **69**(1): p. 7-34.
133. Thompson, I.M., et al., *The influence of finasteride on the development of prostate cancer.* New England journal of medicine, 2003. **349**(3): p. 215-224.
134. Draisma, G., et al., *Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer.* Journal of the National Cancer Institute, 2003. **95**(12): p. 868-878.
135. Etzioni, R., et al., *Overdiagnosis due to prostate-specific antigen screening: lessons from US prostate cancer incidence trends.* Journal of the National Cancer Institute, 2002. **94**(13): p. 981-990.
136. Miller, D.C., et al., *Incidence of initial local therapy among men with lower-risk prostate cancer in the United States.* Journal of the National Cancer Institute, 2006. **98**(16): p. 1134-1141.
137. Telesca, D., R. Etzioni, and R. Gulati, *Estimating lead time and overdiagnosis associated with PSA screening from prostate cancer incidence trends.* Biometrics, 2008. **64**(1): p. 10-19.
138. Etzioni, R., et al., *Quantifying the role of PSA screening in the US prostate cancer mortality decline.* Cancer Causes & Control, 2008. **19**(2): p. 175-181.
139. Ng, M.K., et al., *Prostate-specific antigen (PSA) kinetics in untreated, localized prostate cancer: PSA velocity vs PSA doubling time.* BJU international, 2009. **103**(7): p. 872-876.
140. Wu, Z., et al., *Trajectories of prostate-specific antigen after treatment for prostate cancer.* Journal of Investigative Medicine, 2018. **66**(4): p. 768-772.
141. Moyer, V.A., *Screening for chronic kidney disease: US Preventive Services Task Force recommendation statement.* Annals of internal medicine, 2012. **157**(8): p. 567-570.
142. *About eGFR*. 2019 [cited 2019 4/16]; Available from: https://renal.org/information-resources/the-uk-eckd-guide/about-egfr/.
143. Lenart, M., et al. *Identifying risk of progression for patients with Chronic Kidney Disease using clustering models*. in *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS)*. 2016. IEEE.
144. Medicare, C.f. and M. Services, *Healthcare Common Procedure Coding System (HCPCS)*. 2003: Centers for Medicare & Medicaid Services.
145. Foley, M.M., et al., *Translation Please: Mapping Translates Clinical Data between the Many Languages That Document It.* Journal of AHIMA, 2007. **78**(2): p. 34-38.
146. Zhao, J., et al., *Learning from heterogeneous temporal data in electronic health records.* Journal of biomedical informatics, 2017. **65**: p. 105-119.
147. Zhao, J. and A. Henriksson, *Learning temporal weights of clinical events using variable importance.* BMC medical informatics and decision making, 2016. **16**(2): p. 71.
148. Yuan, Y., et al. *Wave2vec: Learning deep representations for biosignals*. in *2017 IEEE International Conference on Data Mining (ICDM)*. 2017. IEEE.
149. Aghabozorgi, S., A.S. Shirkhorshidi, and T.Y. Wah, *Time-series clustering–A decade review.* Information Systems, 2015. **53**: p. 16-38.
150. Zhao, J., et al. *Detecting adverse drug events with multiple representations of clinical measurements*. in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2014. IEEE.
151. Policker, S. and A.B. Geva, *Nonstationary time series analysis by temporal clustering.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2000. **30**(2): p. 339-343.
152. Aghabozorgi, S., et al., *A hybrid algorithm for clustering of time series data based on affinity search technique.* The Scientific World Journal, 2014. **2014**.
153. Lehman, L., et al. *Similarity-based searching in multi-parameter time series databases*. in *2008 Computers in Cardiology*. 2008. IEEE.

154. Lin, J. and Y. Li. *Finding structurally different medical data*. in *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*. 2009. IEEE.
155. Saria, S., D. Koller, and A. Penn. *Learning individual and population level traits from clinical temporal data*. in *Proceedings of Neural Information Processing Systems*. 2010. Citeseer.
156. Caiado, J., N. Crato, and D. Peña, *Comparison of times series with unequal length in the frequency domain.* Communications in Statistics—Simulation and Computation®, 2009. **38**(3): p. 527-540.
157. Gowda, K.C. and G. Krishna, *Agglomerative clustering using the concept of mutual nearest neighbourhood.* Pattern recognition, 1978. **10**(2): p. 105-112.
158. Park, H.-S. and C.-H. Jun, *A simple and fast algorithm for K-medoids clustering.* Expert systems with applications, 2009. **36**(2): p. 3336-3341.
159. Gordon, T., et al., *High density lipoprotein as a protective factor against coronary heart disease: the Framingham Study.* The American journal of medicine, 1977. **62**(5): p. 707-714.
160. Miller, M., et al., *Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association.* Circulation, 2011. **123**(20): p. 2292-2333.
161. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques.* Emerging artificial intelligence applications in computer engineering, 2007. **160**: p. 3-24.
162. James, G., et al., *An introduction to statistical learning*. Vol. 112. 2013: Springer.
163. Bellman, R., *Curse of dimensionality.* Adaptive control processes: a guided tour. Princeton, NJ, 1961.
164. Bermingham, M.L., et al., *Application of high-dimensional feature selection: evaluation for genomic prediction in man.* Scientific reports, 2015. **5**: p. 10312.
165. Chen, T. and C. Guestrin, *XGBoost: A scalable tree boosting system in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'16 785–794*. 2016, ACM Press.
166. Guyon, I., et al., *Gene selection for cancer classification using support vector machines.* Machine learning, 2002. **46**(1-3): p. 389-422.
167. Yamada, M., et al., *High-dimensional feature selection by feature-wise kernelized lasso.* Neural computation, 2014. **26**(1): p. 185-207.
168. Brodersen, K.H., et al. *The balanced accuracy and its posterior distribution*. in *2010 20th International Conference on Pattern Recognition*. 2010. IEEE.
169. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.
170. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python Journal of Machine Learning Research.* 2011.
171. Jang, T.L., et al. *Low risk prostate cancer in men under age 65: the case for definitive treatment*. in *Urologic Oncology: Seminars and Original Investigations*. 2007. Elsevier.
172. Munro, D. *Annual U.S. Healthcare Spending Hits $3.8 Trillion*. 2014 [cited 2019 05/03]; Available from: https://www.forbes.com/sites/danmunro/2014/02/02/annual-u-s-healthcare-spending-hits-3-8-trillion/#554b035476a9.
173. Buttorff, C., T. Ruder, and M. Bauman, *Multiple chronic conditions in the United States*. 2017: RAND Santa Monica, CA.
174. Ward, B.W., J.S. Schiller, and R.A. Goodman, *Peer reviewed: multiple chronic conditions among us adults: a 2012 update.* Preventing chronic disease, 2014. **11**.
175. Meraya, A.M., A.D. Raval, and U. Sambamoorthi, *Peer reviewed: chronic condition combinations and health care expenditures and out-of-pocket spending burden among adults, Medical Expenditure Panel Survey, 2009 and 2011.* Preventing chronic disease, 2015. **12**.
176. Kulshreshtha, A., et al., *Use of remote monitoring to improve outcomes in patients with heart failure: a pilot trial.* International journal of telemedicine and applications, 2010. **2010**: p. 3.

177. Darkins, A., et al., *Care Coordination/Home Telehealth: the systematic implementation of health informatics, home telehealth, and disease management to support the care of veteran patients with chronic conditions.* Telemedicine and e-Health, 2008. **14**(10): p. 1118-1126.
178. Malasanos, T.H., et al., *Improved access to subspecialist diabetes care by telemedicine: cost savings and care measures in the first two years of the FITE diabetes project.* Journal of telemedicine and telecare, 2005. **11**(1_suppl): p. 74-76.
179. Grabowski, D.C. and A.J. O'Malley, *Use of telemedicine can reduce hospitalizations of nursing home residents and generate savings for medicare.* Health Affairs, 2014. **33**(2): p. 244-250.
180. Dullet, N.W., et al., *Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants.* Value in Health, 2017. **20**(4): p. 542-546.
181. Brewer, R., G. Goble, and P. Guy, *A peach of a telehealth program: Georgia connects rural communities to better healthcare.* Perspectives in Health Information Management/AHIMA, American Health Information Management Association, 2011. **8**(Winter).
182. Boleman, A.B., *Georgia's Telemedicine Laws and Regulations: Protecting Against Health Care Access.* Mercer L. Rev., 2016. **68**: p. 489.
183. Singh, R., et al., *Sustainable rural telehealth innovation: a public health case study.* Health Services Research, 2010. **45**(4): p. 985-1004.
184. Stewart, M.A., *Effective physician-patient communication and health outcomes: a review*. CMAJ: Canadian Medical Association Journal, 1995. **152**(9): p. 1423.
185. Baharav, O., et al., *Distributed system and method for managing communication among healthcare providers, patients and third parties*. 2004, Google Patents.
186. Wang, J.-L., et al., *The role of stress and motivation in problematic smartphone use among college students.* Computers in Human Behavior, 2015. **53**: p. 181-188.
187. Aging, C.f.T.a. *Measuring Return on Investment of Remote Patient Monitoring* 2014.
188. Lee, E.K., et al., *Modeling and optimizing the public-health infrastructure for emergency response.* Interfaces, 2009. **39**(5): p. 476-490.
189. *Counties in Appalachia*. [cited 2019 5/2]; Available from: https://www.arc.gov/appalachian_region/countiesinappalachia.asp.
190. Explorer, S., *Social Explorer*. 2013.
191. Marshall, J., et al., *Health Disparities in Appalachia August 2017 (Creating a Culture of Health in Appalachia: Disparities and Bright Spots). Raleigh, NC: PDA, INC.; Chapel Hill, NC: The Cecil G.* Sheps Center for Health Services Research The University of North Carolina at Chapel Hill.
192. *Places API.* 2019.
193. Dibattista, J. *Is There a Real Time Advantage to Telemedicine?* 2019 [cited 2019 05/03]; Available from: https://www.miramedgs.com/web/63-focus/current-issue/winter-2018/755-is-there-a-real-time-advantage-to-telemedicine.
194. Vogeli, C., et al., *Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs.* Journal of general internal medicine, 2007. **22**(3): p. 391-395.
195. Patel, S., et al., *A review of wearable sensors and systems with application in rehabilitation.* Journal of neuroengineering and rehabilitation, 2012. **9**(1): p. 21.
196. Zur, O., *Reviewing the debate on Skype & HIPAA compliance and introducing the alternative option*. 2016, Retrieved month/day/year from http://www.zurinstitute.com/skype_telehealth
197. Szydło, T. and M. Konieczny, *Mobile devices in the open and universal system for remote patient monitoring.* IFAC-PapersOnLine, 2015. **48**(4): p. 296-301.
198. Fourer, R., D.M. Gay, and B.W. Kernighan, *AMPL: a modeling language for mathematical programming*. Vol. 1. 1993: Scientific Press San Francisco.
199. Budget, O.o.P.a. *Georgia 2030 Population Projections*. 2010.
200. Health, D.o.C. *THE GEORGIA VOLUNTEER HEALTH VOLUNTEER HEALTH CARE PROGRAM*. 2009.

201. McGinnis, J.M., M.H. Lopez, and G. Kaplan, *Transforming health care scheduling and access: Getting to now*. 2015: National Academies Press.